論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Urinary Stones Segmentation by Cascaded U-Net Pipeline and GAN-based Synthetic Stone Augmentation
著者(和文)	PREEDANANWongsakorn
Author(English)	Wongsakorn Preedanan
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12494号, 授与年月日:2023年6月30日, 学位の種別:課程博士, 審査員:熊澤 逸夫,奥村 学,鈴木 賢治,渡辺 義浩,篠﨑 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12494号, Conferred date:2023/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Urinary Stones Segmentation by Cascaded U-Net Pipeline and GAN-based Synthetic Stone Augmentation

by

Wongsakorn Preedanan

Under the supervision of

Prof. Itsuo Kumazawa

東京工業大学
Tokyo Institute of Technology

Information and Communications Engineering
Tokyo Institute of Technology

Acknowledgments

I express my deepest gratitude towards my supervisor Prof. Itsuo Kumazawa for the constant help and encouragement from the starting of the thesis work. I have been fortunate to have a supervisor who gave me the freedom to conduct the research. At the same time supported me and gave suggestions wherever required.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Kenji Suzuki, Prof. Yoshihiro Watanabe, Prof. Manabu Okumura, and Prof. Takahiro Shinozaki not only for their insightful suggestions and encouragement in this research, but also for their supports and opportunities during my Ph.D life for many years.

Furthermore, I am thankful to medical doctors in Department of Urology at Tokyo Medical and Dental University (TMDU) for providing the dataset used in this research and their insightful suggestions to conduct this research.

A big thank you to my special person, Arisa Poonsri, for encouraging and supporting me to conduct this research throughout many years. Finally, I would like to thank my family: Wasan Preedanan, Karnjana Preedanan and Kotchaporn Preedanan, for their love, support and endless encouragement. Without them it might been impossible to finish this work. The author would like to dedicate this work to them.

Tokyo
May 28, 2023

Wongsakorn Preedanan

Contents

List of Tables

List of Figures

Abstract

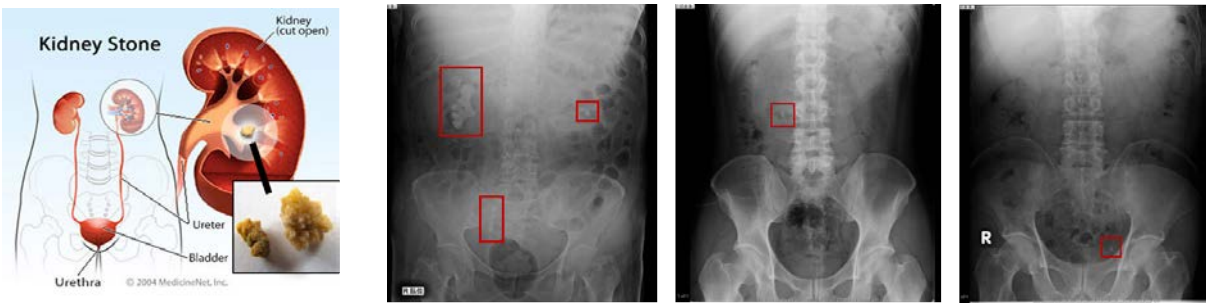
Urinary stones are a common abnormality in the urinary system. Automated segmentation of urinary stones in abdominal x-ray images is demanded to support doctors in screening, diagnosis, and treatment planning. In this research, we propose a two-stage pipeline for segmenting urinary stones. The first stage network generates the KUB (Kidneys,Ureters, and Bladder) region map, representing the approximate locations of urinary organs where stones are present in full abdominal x-ray images. The second stage network takes the partitioned inputs cropped by KUB region maps to generate the segmented stones results.

Recently, deep learning methods have been proposed and played a major role in medical image applications. The performance of deep learning model largely relies on the size of training samples. However, The availability of abdominal x-ray image dataset is limited similar to other medical imaging domains because of difficulty in data acquisition, privacy restrictions, and the need for expert annotators. To address this limitation, we propose a GAN-based synthetic stone augmentation to inpaint the synthetic stones based on the stone masks. This augmentation is utilized with the KUB region maps to create the stone-synthesized images from stone-free images, hence increasing the amount and diversity of training samples. Additionally, the fine-tuned VGG16 mode to classify between real stones and phleboliths, and we utilize the stone location maps from the first stage to reduce false positive results. From the experimental results, the second stage network trained with the combination of real stone-contained images and stone-synthesized images to segment urinary stones can achieved a higher pixel-wise F score and region-wise F score than the baseline method.

Chapter 1

Introduction

1.1 Research Motivation



(a) Urinary Stones (b) Abdominal X-ray Images

Figure 1.1: Urinary stones occurring in urinary organs (a) (image source: [2]), and their appearance in abdominal x-ray images (b) (urinary stones shown in red boxes).

A urinary stone or kidney stone is one of the most common abnormalities in the urinary system. It is a hard deposit made of urine minerals. It usually originates in the kidneys and travels down the urinary tract into the ureters and bladder [1]. Symptoms of a urinary stone include flank pain and gross hematuria in the urine. Each year, more than half a million people visit emergency rooms for urinary stone problems [3]. Therefore, early diagnosis is essential to treat patients promptly before the disease becomes severe [4].

Urinary stones can be detected by using various medical imaging modalities such as CT-scanning, ultrasonography, and x-ray imaging. An abdominal x-ray or KUB (Kidney, Ureter, Bladder) radiography can detect urinary stones well because most stones are calcified. Although radiography is not frequently used for stone detection, advantages of this method include relatively lower radiation exposure than CT imaging and a lower cost than ultrasonography and CT imaging [5]. However, stone detection in plain x-ray images is often difficult for radiologists and other medical doctors because of the following challenges. In radiography, stones and other anatomic structures are projected in a 2D image; hence small stones are difficult to detect because the overlaps, and some types of stones such as irregular one is poorly visible. Computer-aided diagnosis (CAD) is in demand, because it can support radiologists and other medical doctors in various processes, such as screening, diagnosis, and treatment planning. As such, many researchers have proposed approaches to detect or segment urinary stones in ultrasonography [6, 7, 8, 9, 10, 11, 12, 13, 14] or CT-scan imaging [15, 16]. From our best knowledge,

our study is the first research proposing deep learning model for kidney stones detection in abdominal x-ray images (plain KUB radiography).

1.2 Research Challenges

The challenges of our research can be described as follows:

1.) Because plain KUB radiography only views stones at one angle, some stones overlaying bones, or shaded by a bowel gas loop can be poorly visible.

2.) The availability of our plain KUB radiography dataset is limited similar to other medical imaging domains because of difficulty in data acquisition, privacy restrictions, and the need for expert annotators.

3.) In plain KUB radiography, the region of interest (urinary stones) is very small compared with the image background (highly imbalanced). Some small stones can easily be missed.

4.) The majority of urinary stones are found in the kidneys region while stone samples in ureters and the bladder region are rare. Therefore, segmentation performance in these regions can become decreased.

5.) Urinary stones have various sizes and shapes. An abdominal x-ray image usually has multiple stones and area of some stones can be much larger than small ones. Therefore, the large stones are overshadowed the small ones.

1.3 Research Contributions

In this work, we proposed the pipeline of a cascaded framework based on the U-Net deep learning model for the urinary stones segmentation in plain x-ray images. To the best of our knowledge, this is the first study that developed the deep learning techniques to detect the stones from plain abdominal x-ray images. The significant contributions of our work are summarized as follows:

1.) We proposed the pipeline for urinary stone segmentation using two stages of U-Net models. This framework can reduce class imbalance and improve segmentation performance.

2.) We introduced the stone inpainting augmentation by training GAN-based inpainting network to fill the stone-masked region.

3.) We utilized the stone-synthesized images by combining them with the real samples for training the second stage U-Net. This method can increase the number and diversity of training samples.

4.) We modified the training loss function by implementing the stone size re-balancing approach, improving the recall rate of small stones.

Chapter 2

Background and Related Work

2.1 An Overview of Urinary Stones

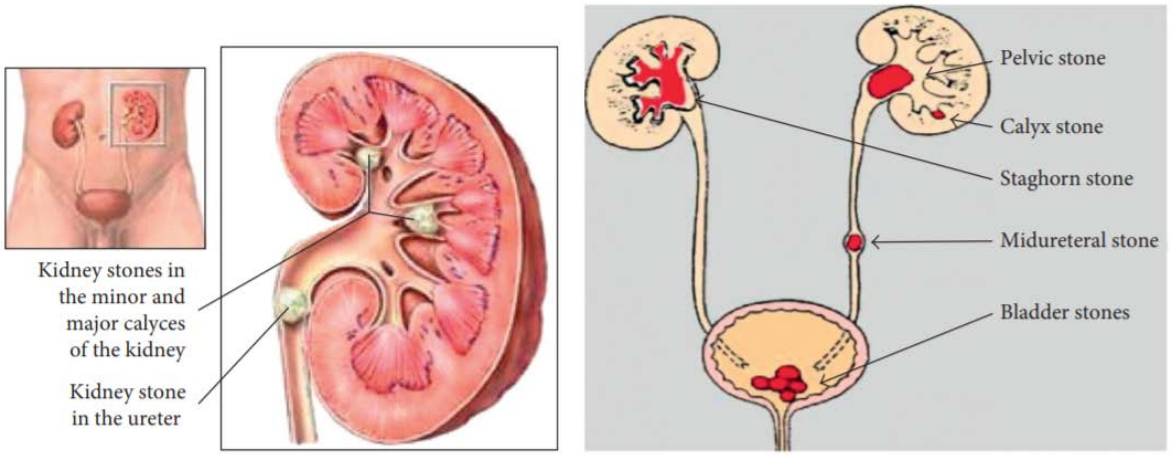
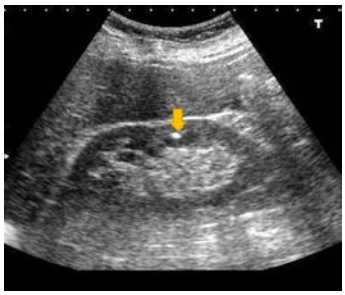


Figure 2.1: Urinary stone locations in the urinary system [4]

Urinary Stone or kidney stone disease is one of the most common abnormalities in urinary tract, affecting about 12% of the world's population [4]. It is a pebble like structure formed by solid concretion by mineral components that originated within kidneys. Symptoms of a urinary stone include lower abdominal pain and gross hematoria in urine. Urinary stones have been related with an increased risk of various diseases such as renal failure, diabetes, and cardiovascular diseases. Urinary stones are located in the kidneys, and only a few of them are lodged in the bladder and urethra as indicated in Figure 2.1. Small stones ($<$ 5 mm) can easily pass the urinary tracts, while the larger stones (5-7 mm) are difficult to pass down the urinary tracts. The remaining stones ($>$ 7 mm) in the kidneys generally have to be surgically removed by invasive or non-invasive techniques.

2.2 An Overview of Urinary Stone Imaging Techniques

(a) CT-scan imaging (KUB position)



(b) Ultrasonography

(c) KUB radiography

Figure 2.2: Medical imaging techniques for urinary stones diagnosis (yellow arrow in each modality points to the urinary stone).

Urinary stone imaging is an important diagnostic method and initial step for doctors to decide the which therapeutic options to use for the urinary stones treatment [5]. The appropriated imaging modality for kidney stones diagnosis involves many factors including the clinical setting, patient body information, cost, and tolerance of radiation dose. The common clinical use includes CT scan imaging, ultrasonography, and kidney ureter bladder (KUB) plain film radiography as displayed in Figure 2.2 (a) - (c), respectively.

2.2.1 CT-Scan Imaging

CT scan imaging at KUB region is an often imaging modality used detecting urinary stones. This modality generates a 3D image of the stone and the surrounding anatomy, which can be reconstructed into multiple views. The sensitivity of CT for detecting stones is the highest of all the available modalities. Limitations of this modality include high cost and radiation exposure.

2.2.2 Ultrasonography

Ultrasonography is a low-cost imaging modality that does not use on radiation and one of the common alternative diagnosis method for urinary stones. This modality makes use of physical differences between stones and surrounding tissues to detect the stones like other imaging modalities, which stone appears brighter than surrounding region. ultrasonography has less sensitive and specific than CT imaging for detecting stones.

2.2.3 Kidney, Ureter, Bladder (KUB) Radiography

KUB plain film radiography or abdominal x-ray imaging uses a single energy source to produce photons, which pass through tissues to the receptor in an anterior-to-posterior orientation to create the image. This modality uses the same fundamental concepts as CT imaging but in a single plane. Advantages of KUB radiography include relatively low radiation dose compared with CT scanning and low cost (10% of ultrasonography). However, as this imaging modality only views stones at one angle, sensitivity and specificity are decreased. Although many stone types can be visualized using this modality, some stone type such as cystine and uric acid stones

often are poorly visible or even not visible at all. Furthermore, stones overlaying bones, or shaded by a bowel gas loop can easily be missed. KUB radiography has a very low sensitivity in small stones (< 5 mm), and the sensitivity increase in the larger stones. Overall, this modality is cost effective compared with other methods for monitoring stone size in a patient with known stone disease and received medical therapy.

2.2.4 Summary

CT scanning is the most accurate imaging modality for urinary stones diagnosis because of high sensitivity, specificity, accurate stone sizing, and the ability to evaluate other pathology. Ultrasonography has a lower sensitivity and specificity than CT, but does not produce ionizing radiation and is less expensive than CT. An abdominal x-ray or KUB (Kidney, Ureter, Bladder) radiography is not frequently used for stone detection compared to those methods because of many limitations. In radiography, stones and other anatomic structures are projected in a 2D image, so small stones are difficult to detect because the overlaps, and some types of stones such as irregular one is poorly visible, which is often difficult for radiologists or even experts. The comparison between imaging modalities for urinary stones diagnosis is shown in Table 2.1.

Table 2.1: The comparison of different imaging modalities for urinary stones diagnosis.

	CT scan	Ultrasonography	Plain x-ray imaging
Radiation dose	High	-	Low
Cost	High	Medium	Low
Accuracy (by doctors)	High	Medium	Low

2.3 Computer Aided Diagnosis for Urinary Stones Detection Works

Computer-aided diagnosis (CAD) for urinary stones detection is demanded for supporting radiologists and medical doctors in various processes, such as screening, diagnosis, and treatment planning. Several works have been proposed kidney stone detection in ultrasound images using various segmentation approaches [6, 7, 8, 9, 10, 11, 12, 13, 14], such as region growing [6], contour-based square Euclidean distance [8], and region indicator with contour segmentation method [11]. The recent work in [14] extracted texture-based features from candidate segmented stone regions, and used KNN classifier to classify them. For the stone detection in CT-scan images, the authors in [15] proposed the level set segmentation method to detect kidney stones. The work in [16] used intensity-based and sized-based thresholding to remove unwanted objects to detect the stones, and used feature-based thresholding to reduce false positive results.

2.4 U-Net Based Deep Learning Models

2.4.1 U-Net

The U-Net architecture was proposed by O. Ronneberger et al. [17]. It is the encoder-decoder deep learning architecture built based on fully convolutional neural network. Its architecture is separated in 2 parts: contracting path (encoder) and expanding path (decoder) as illustrated in Fig.2.3. Skip connection between encoder and decoder path is implemented to combine the

location information from the encoder path with the contextual information in the decoder path to obtain a general information combining localization and context, which is necessary to predict a good segmentation map [18]. It has been used in many medical image segmentation task and achieved very good performance. It has been proved that it can work effectively with a few training images and has a very reasonable training time.

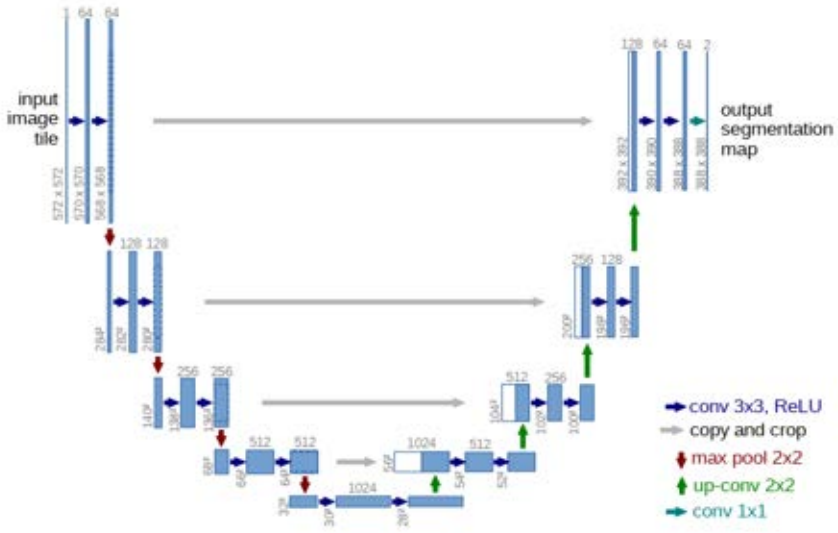
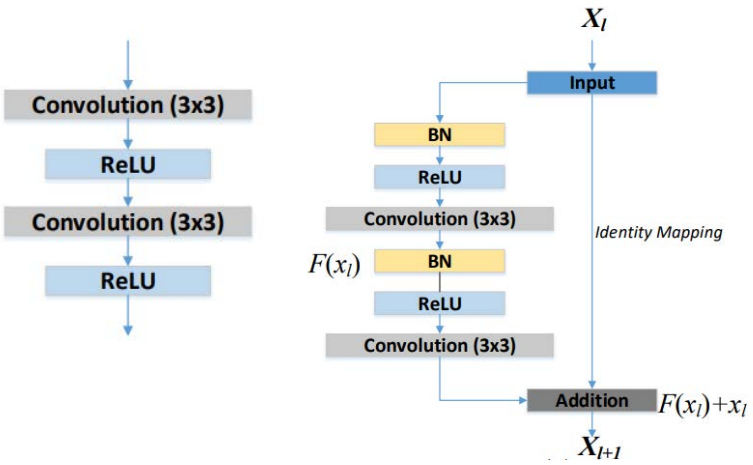


Figure 2.3: U-Net architecture proposed by O. Ronneberger et al. 2015.

2.4.2 Residual U-Net

Deep neural network would have better performance when the number of convolutional layers is increased. However, it is very difficult to train a very deep architecture due to vanishing gradients problem. He et al. proposed the deep residual learning framework that utilize an identity mapping to facilitate training deep networks. Therefore, in the segmentation task, Z. Zhang et al. [19] proposed the deep residual U-Net, an architecture that utilize of strengths from both deep residual learning and U-Net architecture. This network use residual units instead of plain convolutional units as basic blocks to build the deep ResUnet as shown in Fig. 2.4.



(a) (b)

Figure 2.4: Building blocks of neural networks. (a) convolution block used in original U-Net model and (b) residual unit with identity mapping used in the ResUnet.

2.4.3 U-Net ++

Although, U-Net model have been proved to work impressively in semantic segmentation tasks. Lesions or abnormalities segmentation in medical images demands a higher level of accuracy than natural images. While a precise segmentation mask might not be as critical in natural images as medical images. Even marginal segmentation errors in medical images can lead to poor user experience in clinical settings. Fine details of lesion or abnormalities are crucial information for medical doctors. Therefore, Z. Zhou et al. [20] proposed UNet++, a new encoder-decoder segmentation architecture based on nested and dense skip connections to improve U-Net-based model. In the Fig 2.5, convolutional units in black color are the original U-Net, green and blue show dense convolution blocks on the skip pathways, and red indicates deep supervision. Red, green, and blue components distinguish UNet++ from original U-Net.

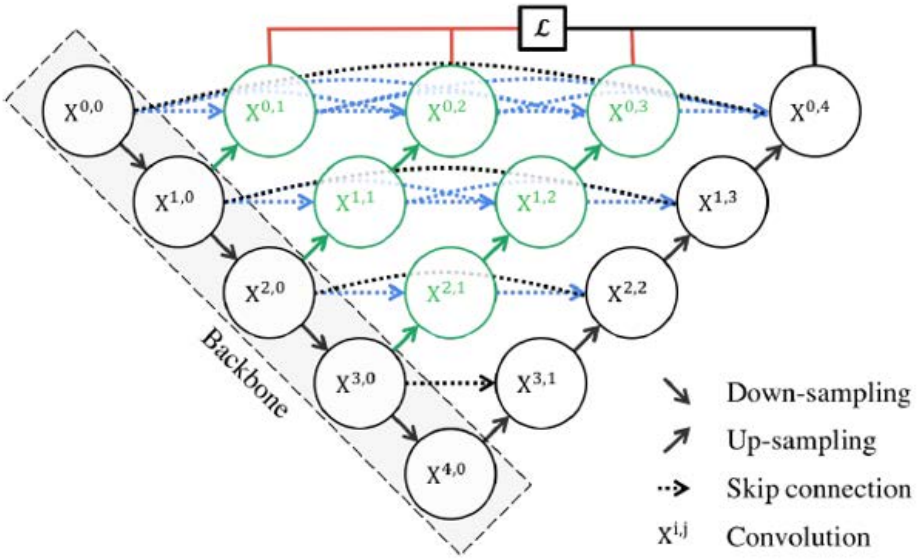
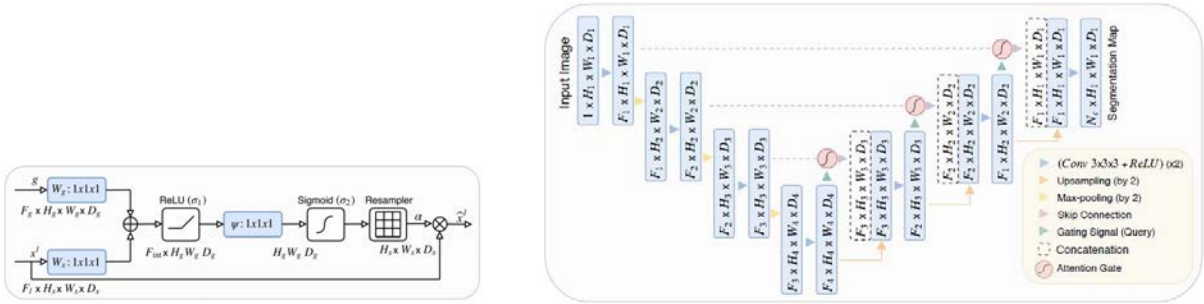


Figure 2.5: UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks.

2.4.4 Attention Unet



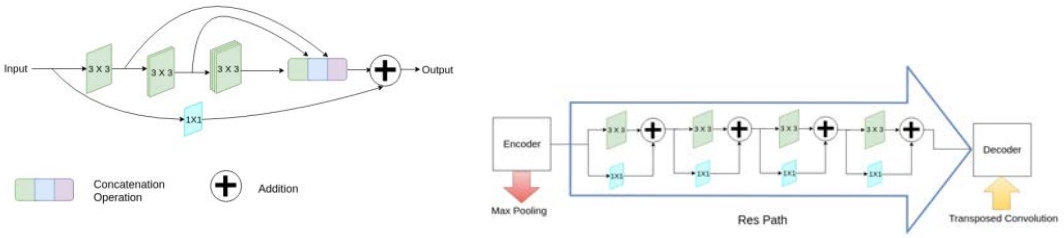
(a) (b)

Figure 2.6: The proposed Attention U-Net segmentation model (a) and schematic of the proposed additive attention gate(AG) (b).

The authors in this work [21] proposed to use attention gate (AG) structure for U-Net model to automatically lean to focus on the target structures that have different shapes and sizes. This attention gate structure can be described in Fig.2.6 (a), and integrated with U-Net model in the skip connections at each level as shown in Fig.2.6 (b). Instead of combining the contextual information from the deeper level of convolutional layers with the spatial information from the skip connections at the shallow layers like the standard U-Net model, the proposed AG mechanism takes two inputs, skip connections from encoder and the layer from decoder representing better feature representation. The two vectors are summed element-wise. Then, the resultant vector goes through a ReLU activation, and sigmoid layer respectively. Therefore, this process results in aligned weights becoming larger while unaligned weights become relatively smaller. The networks trained with this mechanism can learn to suppress irrelevant regions in an input image while highlighting salient features that useful for a specific task.

2.4.5 MultiRes U-Net

N. Ibtehaz et al. [22] proposed MultiResUnet model, which is the modifications of the U-Net-based model for medical image segmentation. The MultiResUnet replacs the convolutional layers with Inception-like blocks called MultiRes blocks to reconcile the features learnt from the image at different scales, which is a common problem for lesion segmentation tasks. MultiRes block utilizes the succession of 3×3 filters instead of using the multi-resolution filters (3×3, 5×5 and 7×7 filters) as shown in Fig. 2.6 (a). Furthermore, instead of the plain skip connections, MultiResUnet uses the stacks of convolutional unit with residual connection as shown in Fig. 2.6(b). This Res path can reduce the gap between the encoder and decoder path, especially lower level features in the first level of encoder and much more higher level in the last layer of decoder. MultiResUNet architecture is illustrated in Fig. 2.7.



(a) (b)

Figure 2.7: Building blocks of neural networks. (a) multiRes block and (b) Res path used in the MultiResUnet model.

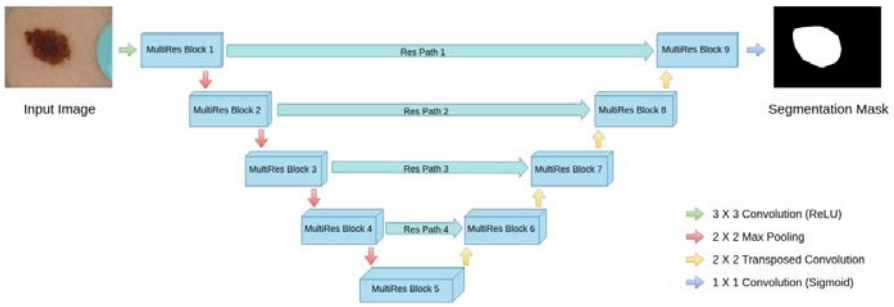


Figure 2.8: MultiResUNet architecture replaces the sequences of two convolutional layers in original U-Net architectures with the MultiRes block, and uses Res path instead of using plain shortcut connections.

2.5 Transformer-based Segmentation Models

The Transformer is a model proposed in the paper "Attention Is All You Need" by Vaswani et al. (2017) [23]. It is a model that uses a mechanism called self-attention and has been widely used in NLP field. Vision Transformer (ViT) is a model that applies the Transformer to the image classification task and was proposed in 2020 by Dosovitskiy et al [24].

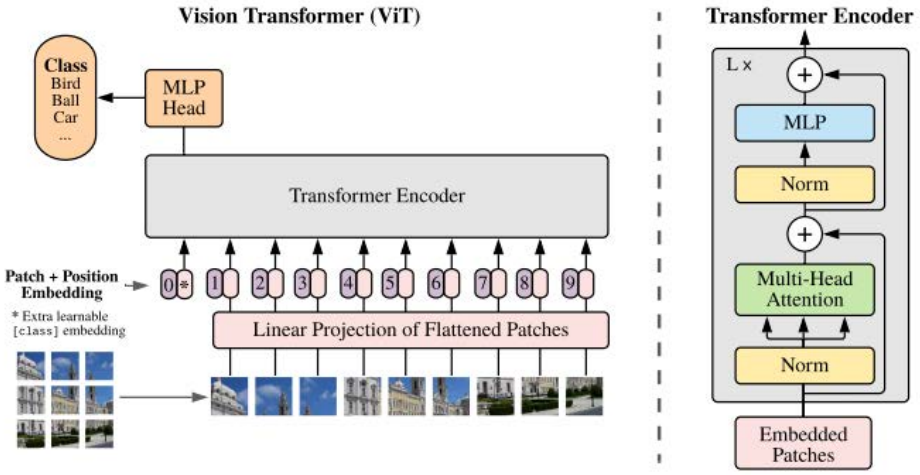


Figure 2.9: An overview of Vision Transformer (ViT). An image is split into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder.

An overview of Transformer Network is described in the Fig. 2.9 An image is split into fixed-size patches, each of them are then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder. Transformer learns by measuring the relationship between input token pairs.

2.5.1 TransUnet

TransUnet [25] was proposed to combine the strong merits of both Transformer model, and Unet model for medical image segmentation task. This model adds the Transformer layers in the encoder path of U-Net model. Transformer model encodes tokenized image patches from the feature maps from convolutional neural network (CNN) as the input sequence for extracting global contexts. While the decoder upsamples the encoded features, which are combined with the high-resolution CNN feature maps in the same spatial dimension to enhance the finer details. The overview architecture of TransUnet is shown in 2.10.

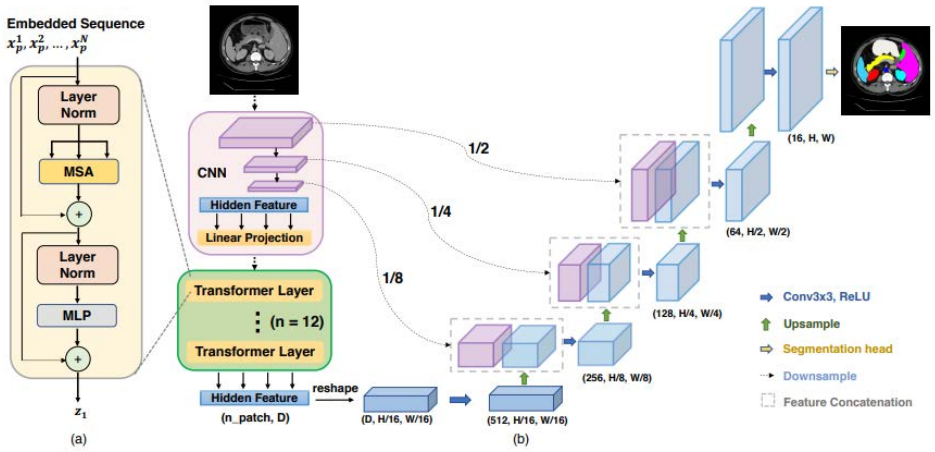


Figure 2.10: Overview of the TransUnet framework. (a) schematic of the Transformer layer; (b) architecture of the proposed TransUNet.

2.5.2 UTNet

The U-shape hybrid Transformer network (UTNet) [26] combined the merits of both convolutional neural network (CNN) and self-attention mechanism for medical image segmentation. The objective of this model is to apply CNN layers to extract the local intensity features to avoid large-scale pre-training of Transformer, while using self-attention to capture long-range relative information. This main architecture of this model is simple and similar design with the Unet-based models, but replace the last layer of each block with the Transformer module to enhance the quality of segmentation results.

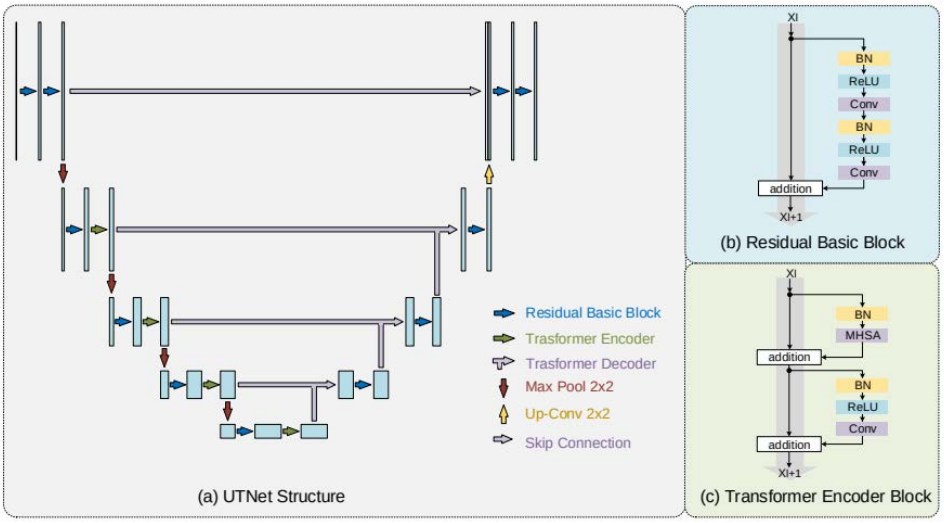


Figure 2.11: Overview of the UTNet framework. (a) schematic of the Transformer layer; (b) architecture of the proposed TransUNet.

2.5.3 Swin Unet

The existing medical image segmentation methods mainly rely on convolutional neural network (CNN) with U-shape architecture. Although the CNN-based approaches have achieved excellent performance inn this field, the segmentation is still a challenging task in medical image application. CNN cannot learn global and long-range semantic information due to the limitation of locality of convolution operation. Therefore, the recent proposed methods tried to combine the Transformer module with the CNN-based backbone to solve this limitation.

Swin U-Net [27] is a Unet-like pure Transformer for medical image segmentation task. The tokenized patches are fed into the Transformer-based U-shaped Encoder-Decoder architecture with skip-connections for local-global semantic feature learning. The hierarchical Swin Transformer with shifted windows is used in the encoder to extract context features. A symmetric Swin Transformer-based decoder with patch expanding layer is used for up-sampling operation to restore the spatial resolution of the feature maps.

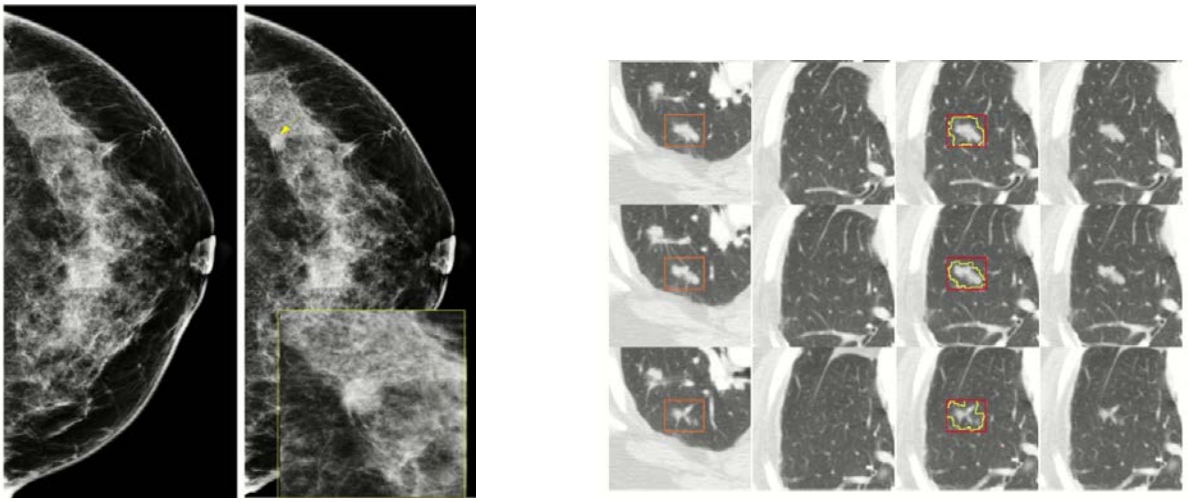
2.6 Lesions Augmentation in Medical Imaging

Deep learning has been widely used in various medical imaging tasks and shown improvements over traditional feature engineering methods [28]. However, the performance of deep learning usually depends on the number of training data. The availability of medical image datasets is

limited compared with other domains because of the data acquisition cost, privacy restriction, and difficulties in image labelling, which need experts. Furthermore, the class imbalance is also the common problem in medical domains, which the normal samples are dramatically out number the samples with lesion. Augmentation of small or imbalanced training datasets by synthetically generating additional training data has been proved to provide the classifier with a better representation of data for those classes with a small number of samples, which can improve the generalization of the networks. Basic data augmentation techniques, such as image shifting, scaling, flipping and rotations, usually be used to increase data diversity during the training stage, but cannot be used for the diversity of nodule characteristics and locations. Accordingly, many studies proposed the various methods to synthesize new positive training samples.

2.6.1 Traditional Techniques for Lesion Insertion on Medical Images

In medical imaging application, many techniques for creating new synthetic samples and varying the distribution of lesion properties in an existing real dataset have been developed. The past works for lesions insertion on medical images can be divided into 2 categories. In the first category, the new lesions are simulated using a mathematical model and inserted into the existing medical images using various blending approaches. The study in [29] simulated lung nodules that have realistic characteristics in inserting them into CT slides. The one in [30] utilized a physic-based method for simulating 3D lesions and projected them in 2D on mammogram images. And the one in [31] generating 3D lesions from real 2D lesions extracted from digital breast tomosynthesis (DBT), and inserting them to healthy samples. In the other category, an actual lesion is extracted from real images and then inserted into the new location on other images using various blending techniques such as the studies in [32, 33, 34, 35]. The recent work in [35] proposed to use the lesion blending approach based on Poission method. The authors used this tool to create the augmented samples, and use them to improve the performance of classifier for small training datasets.



(a) (b)

Figure 2.12: An example of lesion insertion on mammography in [30], which the first image showing the original image and second image showing the result, and an example of lesion insertion on CT imaging (b) [35], which the 1^{st} column showing the source lesion cropped on CT slides, the 2^{nd} column showing the target CT slides, and the 3^{rd} - the 4^{th} columns showing blending process and the final blended outputs, respectively.

2.6.2 Generative Adversarial Networks and the Applications on Synthetic Lesion Augmentation

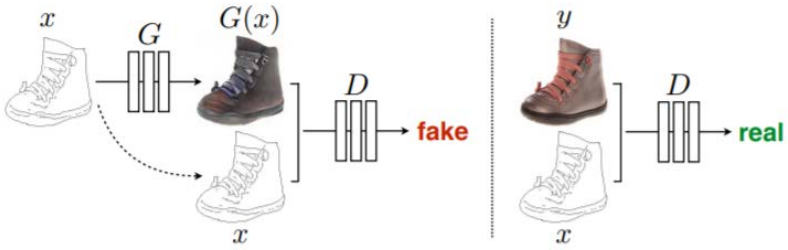
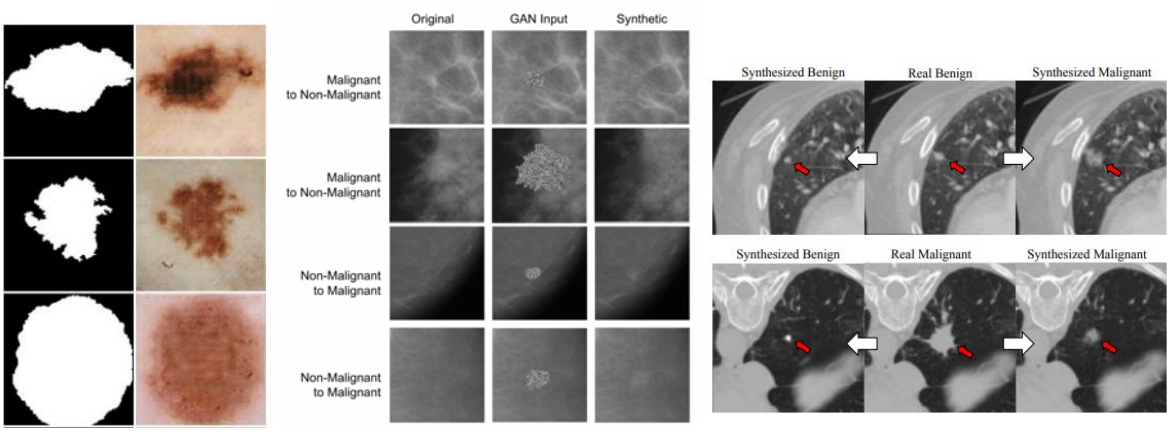


Figure 2.13: A conditional GAN is trained to map an image from one domain (edges) to another domain (photo). The discriminator, D, learns to classify between fake (synthesized image by the generator) and real images with the input pairs. The generator, G, learns to fool the discriminator. (Image source: [37])

I. Good fellow introduced Generative Adversarial Networks (GANs) in 2014 [36]. GANs can be described as the competition between two networks The first network, the generator G, takes the random noise as the input and generate the the synthetic outputs, while the second network, the discriminator D, is the binary classifier, trying to distinguish between real training samples and fake synthetic samples from the generator. Two networks are trained simultaneously, which the generator is trained to maximize the the probability of fooling the classifier (fooling the discriminator to think that the synthetic samples are real). On the other hand, the discriminator is trained to minimize the classification loss between real and generated samples, or maximize the probability of classifying real and synthetic images correctly. Pix2Pix GAN [37] was introduced in 2016 as the general framework for image-to-image translation problems. A conditional GAN learns to map images from one domain to another domain as shown in Fig. 2.9. This work also combines the pixel-wise reconstruction error (L1 loss) with the adversarial loss from the discriminator to train the encoder-decoder generator.

Because of the successful results of GANs in many domains, generative adversarial networks (GANs) have been proposed to use in medical imaging augmentation applications in various applications such as medical image modalities translation in [38, 39, 40], and image denoising in[41, 42]. Furthermore, GANs also utilized for generating the synthetic samples. For examples, the work in [43] proposed to use GANs to generate synthetic liver lesion images for 3 categories of lesion including cyst, metastasis, and hemangioma. The authors used them and the real samples to train CNN classifier, and the results showed an improvement of lesion classification performance. Skin lesion augmentation is another popular researches, which many studies in [44, 45, 46, 47] applied GANs to synthesize new skin lesion images for generating the synthetic samples. The one in [47] utilized the lesion mask dataset to train the generative network to map from the input masks to skin lesion images (Fig. 2.10(a), then used the synthetic samples to improve the performance of lesion segmentation network.



(a) (b)

(c)

Figure 2.14: Illustration of the applications of GANs in Medical Imaging such as the skin lesion augmentation (a), lesion inpainting on mammography (b), lesion inpainting on CT images (c), and region inpainting on brain MRI images (d).

Image inpainting is a task of reconstructing the missing or distorted region in an image. Recently, GANs have been widely used in this application instead of the traditional approaches. Context Encoder (CE) [48] is an auto-encoder architecture training with adversarial loss and reconstruction loss. The studies in [49, 50] improve the CE framework by using two discriminator networks consisting of local discriminator taking the completed region as input and global discriminator taking the entire image as input. More recently, ip-MedGAN [51] is the inpainting framework developing for medical imaging. This method uses cascaded multiple U-Net networks as the generator trained with the combination loss of discriminator networks, reconstruction loss, perception loss, and style loss.

In medical imaging applications, many recent studies proposed to use GANs in image inpainting task to synthesize lesions in medical image patches to augment the training data in mammogram [52], and lung nodule in CT images [53, 54, 55, 56]. These methods train GANs to inpaint a cropped region with the objects of interests, such as lesions. Deep learning trained with synthesized samples has been shown to improve the performance in classification and segmentation tasks.

2.7 Loss Functions for Segmentation

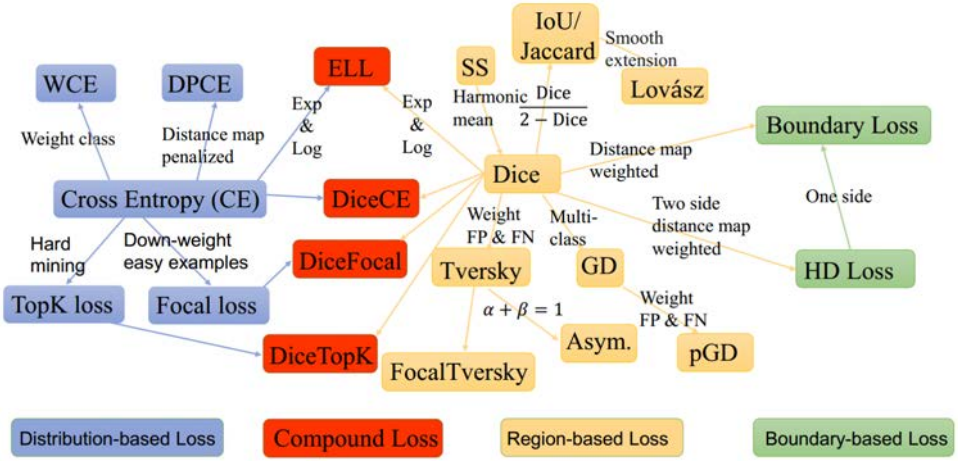


Figure 2.15: The overview of loss functions for image segmentation [57].

Deep Learning methods use stochastic gradient descent approach to optimize objective functions. To obtain the good deep learning model, it is important to define the suitable objective function for the application. loss functions for semantic segmentation can be categorized into 4 categories: Distribution-based, Region-based, Boundary-based, and Compounded loss as shown in Fig. 2.12.

2.7.1 Distribution-based losses

Distribution-based loss functions aim to minimize dissimilarity between two distributions. Cross entropy loss is the most fundamental loss for this category, and other functions are modified from this loss.

Binary Cross-Entropy

Cross-entropy is the most widely used loss for classification objective, and pixel-level classification in the segmentation task. It is a measure of the difference between two probability distributions, which is defined as:

$$L_{BCE}(y, \hat{y}) = -y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \tag{2.1}$$

where y and are the true value and predicted value by the model, respectively.

Weighted Cross-Entropy

Weighted cross-entropy is the variant of binary cross entropy loss. β value in the formula can be used to tune the balance between false negatives and false positives. Weighted cross-entropy can be defined as:

$$L_{WCE}(y, \hat{y}) = -(\beta * y \log(\hat{y})) + (1 - y) \log(1 - \hat{y}) \tag{2.2}$$

Focal loss

Focal loss (FL) [58] is the variant of cross entropy loss that down-weights the contribution of easy examples and enables the model to focus more on learning hard examples. It works well for highly imbalance class tasks. Let p_t defined as:

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise} \end{cases} \tag{2.3}$$

So, Focal Loss (FL) can be written as

$$FL(p_t) = \alpha_t (1 - p_t)^y \log(pt) \tag{2.4}$$

Where γ is the focusing parameter for adjusting the rate at which easy examples are down-weighted as shown in Fig.. When $\gamma = 0$, FL works like Cross-Entropy loss function, which has α_t value as a weight between two classes. $\gamma = 2$ work best for their experiments.

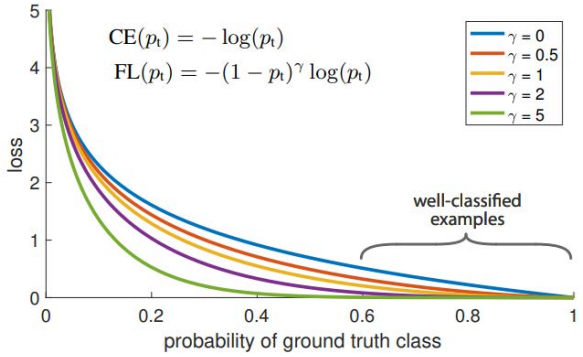


Figure 2.16: The Focal Loss adding a factor $(1pt)^\gamma$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples, putting more focus on hard examples (misclassified examples).

2.7.2 Region-based losses

Region-based loss functions aim to maximize the overlap regions between ground truth g and predicted segmentation p.

Dice Loss

The Dice coefficient is the fundamental objective function for region-based loss. This loss is widely used metric in many medical image segmentation tasks. Dice Loss (DL)

$$DL(y, p) = 1 - \frac{2yp + 1}{y + p + 1} \tag{2.5}$$

Some work used the squared terms in the denominator, which is defined by

$$DL(y, p) = 1 - \frac{2yp + 1}{y^2 + p^2 + 1} \tag{2.6}$$

where 1 is added in numerator and denominator term in both formula to avoid division by zero.

Tversky Loss

Tversky index (TI) [59] is an generalization of Dices coefficient, which adds a weight β to FP (false positives) and FN (false negatives).

$$TI(y, p, \beta) = \frac{yp + 1}{yp + \beta(1 - g)p + (1 - \beta)g(1 - p) + 1} \tag{2.7}$$

when $\beta = 0.5$, It can be used as the regular Dice coefficient. Tversky loss (TL) can also be defined as:

$$TL = 1 - TI \tag{2.8}$$

Focal Tversky Loss

Similar to Focal Loss, which focuses on hard examples by down-weighting easy ones. Focal Tversky loss [60] also aims to learn hard examples such as with small ROIs(region of interest) with γ coefficient, which is defined as:

$$FTL = (1 - TI)^{\gamma} \tag{2.9}$$

where γ ranges from [1,3].

2.7.3 Boundary-based losses

Shape-aware Loss

Loss functions are usually used to evaluate the pixel-wise similarity in whole image, which lacks the shape information. Shape-aware loss [**?**] also calculates the average point to curve Euclidean distance among points around curve of predicted segmentation to the ground truth. It can be defined as:

$$E_i = D(P, G) \tag{2.10}$$

$$L_{shape-aware} = -CE(p, g) - E_i CE(p, g) \tag{2.11}$$

2.7.4 Combined losses

Combo Loss (Dice Cross-Entropy Loss)

Combo loss (CL) [**?**] is defined as a weighted sum of Dice loss and a modified cross entropy. It attempts to leverage the flexibility of Dice loss of class imbalance and at same time use cross-entropy for curve smoothing. It's defined as:

$$CL(y, p) = \alpha L_{BCE} - (1 - \alpha)DL(y, p) \tag{2.12}$$

2.8 Loss Reweighting Approaches in Lesion Size Imbalance

Several studies proposed the reweighting methods to solve the class imbalance problem (lesion pixels vs. non-lesion pixels or multi-classes problem). The common reweighting methods includes Weighted Cross-entropy loss (WCE) and general Dice coefficient loss (GDL). These methods add the different weight to each class to solve class imbalance problem as shown in Fig. 2.13 (b). The recent work in [62] proposed a loss reweighting approach to increase the performance of the network to detect small lesions. This method assigned weights are inversely proportional to the lesion volume, the weight for every components $L_0, \ . \ . \ . \ , L_K$ are assigned by Eq. (2.13), where L_0 is the non-lesion component (background) and K is the number of lesions in the current patch.

$$w_j = \frac{\sum_{k=0}^{K} L_k}{(K+1)\dot{L}_j} \tag{2.13}$$

where w_j is the weight, assigned to every voxel inside the corresponding component L_j, therefore, smaller lesions get larger weights as shown in Fig. 2.13 (c).

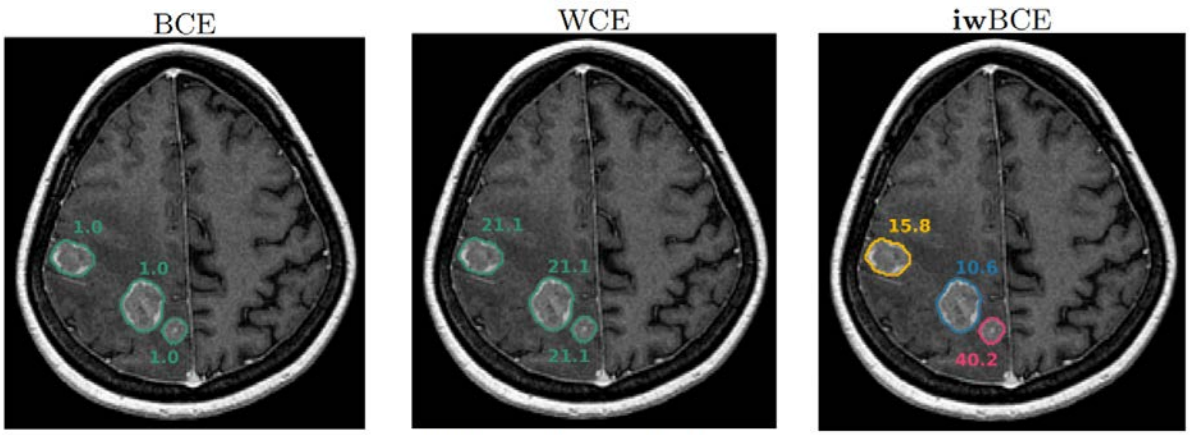


Figure 2.17: The effect of inverse weighting. No reweighting applied (a), class balancing via Weighted-Cross Entropy (b), inverse weighting (c). Weights for every tumor are placed near the tumors.

Chapter 3

Proposed Method

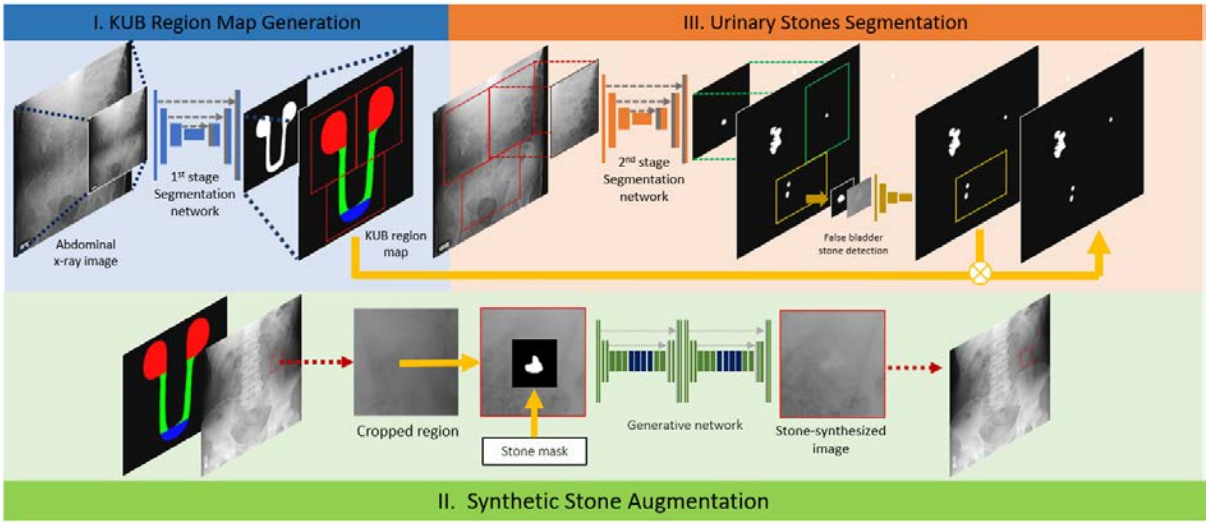


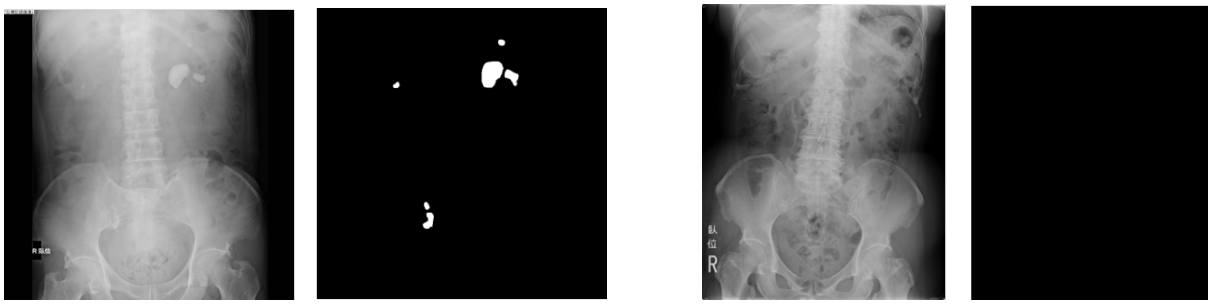
Figure 3.1: The overview of the proposed pipeline for segmenting urinary stones.

Our pipeline has two stages: the KUB region map generation and stone segmentation. The first stage receives a full abdominal x-ray image as input and generates the coarse map representing the region where stones can be found, including kidneys, ureters, and bladder. The outputs from this stage are up-sampled and used for cropping the full image into three smaller partitions, including the right kidney, left kidney, and bladder region, which are used as the input in the second stage and also used as a KUB region map in stone-synthesized augmentation. The proposed augmentation is the augmentation method which synthesizes new stone(s) in the new locations to increase the number and variety of training data. The second stage U-Net was trained with both real samples and synthetic samples to generate the prediction map of urinary stones. The section 3.1 will explain the abdominal x-ray dataset that we used. The first stage work will be described in 3.2. The proposed stone augmentation approaches will be described in 3.3. Lastly, the urinary stones segmentation task will be described in 3.4.

3.1 Abdominal X-ray Images Dataset

The abdominal x-ray images and their gold standard stone masks were provided by Tokyo Medical and Dental University. There are two types of dataset including stone-contained samples

(I_{sc}) from patients who have urinary stones, and stone-free samples (I_{sf}) from healthy person. The stone masks which require medical knowledge and precise annotation skills, were manually drawn by the urology experts for every stone-contained image. The total number of stone-contained samples (I_{sc}) and stone-free samples (I_{sf}) are 1,156 and 1,199 images, respectively.

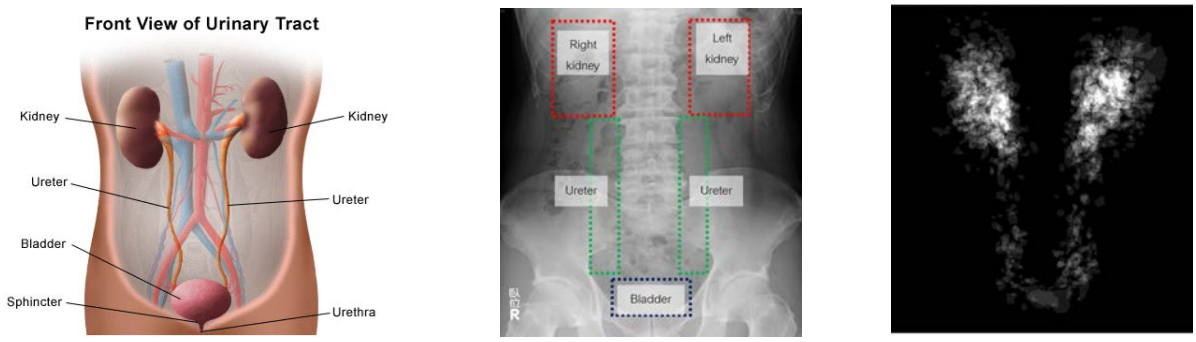


(a) (b)

Figure 3.2: Example of a stone-contained sample (I_{sc}) and its corresponding gold standard manual segmentation of the stones (a) and a stone-free sample (I_{sf}) (b).

3.2 KUB Region Map Generation

3.2.1 Stone Location Map



(a) (b) (c)

Figure 3.3: Illustration displaying anatomy of urinary organs (a) adopted from [2] , the approximated location of urinary organs in an abdominal x-ray image (b), and distribution of stones in our dataset (c)

Based on medical domain knowledge, urinary stones are formed in kidneys and passed down into ureters and bladder. Therefore, they can be only found in urinary organs including kidneys, ureters, and bladder (Fig.3.3 (a)). The approximate locations of which abdominal x-ray images is shown in Fig.3.3 (b). The segmentation map representing the urinary organs in KUB radiography is important for many utilities, however it is difficult to generate the precise segmentation map of these organs because of the following challenges:

1.) The precise boundaries of urinary organs are difficult to be seen in KUB radiography.

2.) The characteristics of KUB radiography (such as brightness, position) in our dataset are varied, so one stone location map cannot be used for all images.

When we plotted all stone from the stone mask dataset, the distribution of urinary stones is displayed as Fig. 3.3(c). The highest density area is shown in kidneys region, and the density decreases in ureters and bladder, respectively. In this task, we can create the ground-truth of stone location map which it shape is U-shape similar to the stone distribution map.

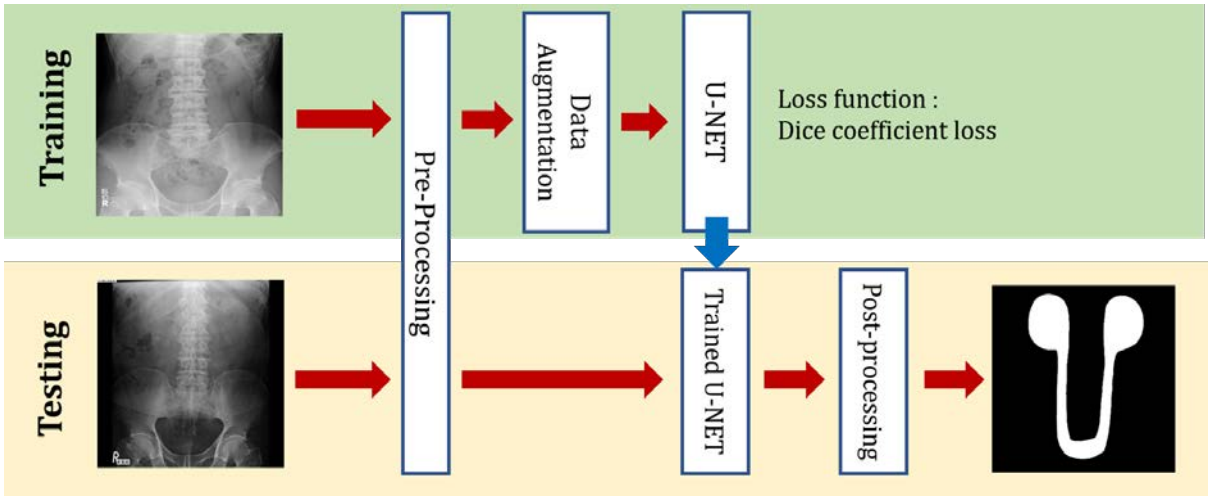


Figure 3.4: The overview of training and testing methodology for stone location map generation

The overview process of training and testing methodology for stone location map generation is illustrated in Fig. 3.4. Firstly, all training images were resized to 256×256 pixels and normalized to zero mean and unit variance. We used the slightly modified U-Net model for training a network to generate coarse stone location maps. The U-Net structure consists of encoder paths and decoder paths that have skip connections in each corresponding layer with the same spatial dimension. In the encoder path, two convolution blocks consisting of a 3×3 convolution layer, batch normalization [63], and leaky ReLU activation function are used in each level. A 3×3 max-pooling with the stride of 2 is implemented to halve the resolution while the number of features channels is twice increased in the following level. In the decoder path, a 2×2 transposed convolution is implemented to twice upsampling the feature map while the number of feature channels is halved. The output from the upsampling is concatenated with the corresponding feature map from the decoder path, followed by a 3×3 convolution layer, ReLU activation, and 0.5 dropouts. In the output layer, a 3×3 convolution layer followed by the Sigmoid activation function is applied for generating the continuous values between 0 and 1, representing the probability of stone location map pixels. The simple image augmentation, including image rotation [-5, 5] and horizontal flipping, were randomly implemented during the training process. The network was trained with 600 images from scratch for 100 epochs and used Adam optimize [64] with a learning rate of 10^{-3} to minimize the Dice coefficient loss (DL) written in Eq.(3.1).

$$DL = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{3.1}$$

where the sums run over the N pixels, of the predicted segmentation pixel p_i and the ground truth binary pixel g_i.

In post-processing, the output images were binarized using a 0.5 threshold value and implemented morphological operations to connect all white components and remove the small ones. Example of stone location map results from this network are shown in Fig. 3.5.

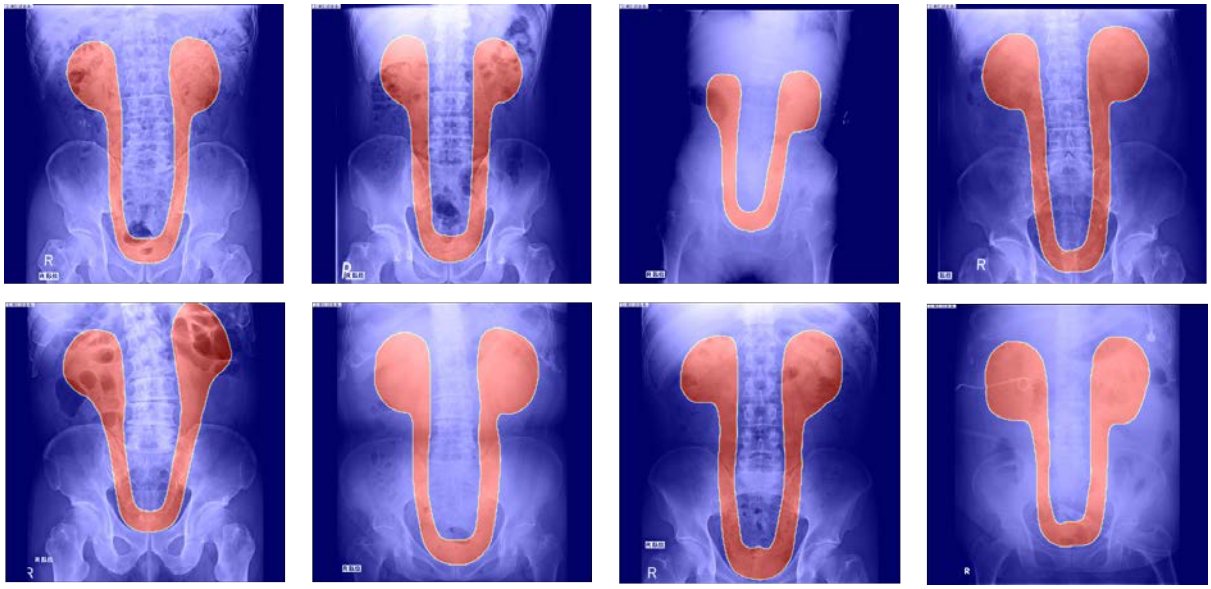


Figure 3.5: The illustration of example stone location map results visualized in heatmap.

3.2.2 KUB Region Map

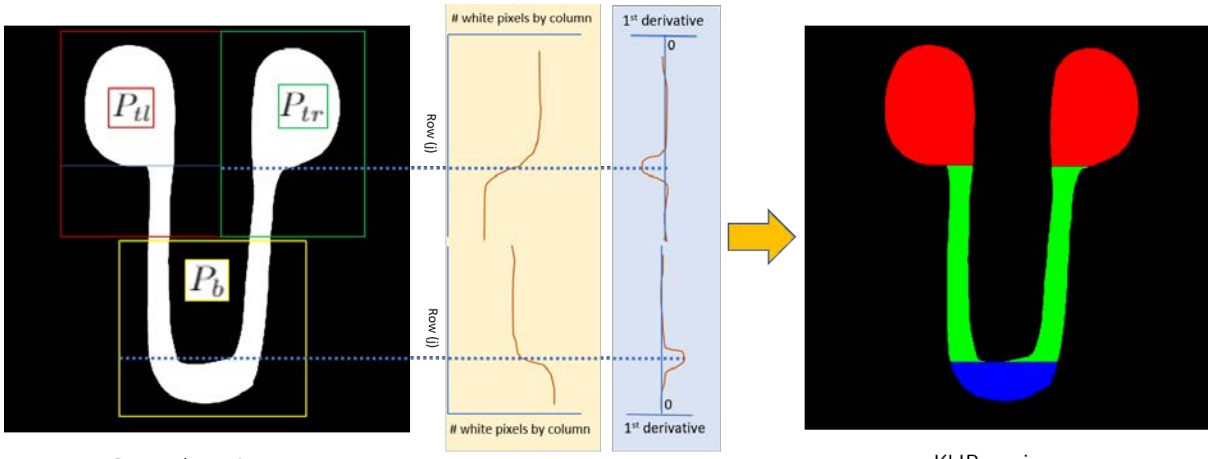


Figure 3.6: The overview process to create KUB region map from stone location map.

In this process, we used stone location maps to create KUB region maps representing kidneys, ureters, and bladder region. The overview process is shown in Fig. 3.6. Let (x^m_{tl}, y^m_{tl}) be the top-left coordinate and (w^m, h^m) be its width and height of the bounding box of a stone location map. We cropped this bounding box into 3 partitions; top-left partition(P_{tl}), top-right partition(P_{tr}), and bottom partition (P_b). The coordinates of top-left partition (x^p_l, y^p_l), top-right partition (x^p_r, y^p_r), and bottom partition (x^p_b, y^p_b) can be defined as Eqs.(3.2) - (3.4), respectively. The width and height of each partition (w^p, h^p) are defined as Eq. (3.5).

$$(x^p_l, y^p_l) = (x^m_{tl} - b_x, y^m_{tl} - b_y) \tag{3.2}$$
$$(x^p_r, y^p_r) = (x^m_{tl} + w^m/2, y^m_{tl} - b_y) \tag{3.3}$$
$$(x^p_b, y^p_b) = (x^m_{tl} + w^m/4, y^m_{tl} + h^m/2 + b_y) \tag{3.4}$$
$$(w^p, h^p) = (w^m/2 + b_x, h^m/2 + b_y) \tag{3.5}$$

where b_x is the border size in the vertical direction and b_y is the border size in the horizontal direction. We set these values equal to 10% of the width and height of the stone location map's bounding box, respectively.

Then, we can split the stone location map into kidneys, ureters, and bladder regions. S_{tl} and S_{tr} which are the region separating lines used in top-left partition (P_{tl}) and top-right partition (P_{tr}) are defined at the row (j) that has the lowest derivative value of the sum of column pixels (i) in a binary stone location image described in Eqs. (3.6) and (3.7), respectively, while the separating line S_b used in bottom partition (P_b) is defined at the row (j) that has the highest derivative value of the sum of column pixels (i) in a stone location image described in Eq. (3.8).

$$S_{tl} =_j \Delta_j(\sum_{i=0}^{w^p} P_{tl}(i,j)) \tag{3.6}$$

$$S_{tr} =_j \Delta_j(\sum_{i=0}^{w^p} P_{tr}(i,j)) \tag{3.7}$$

$$S_b =_j \Delta_j(\sum_{i=0}^{w^p} P_b(i,j)) \tag{3.8}$$

Examples of final KUB region map results are shown in Fig. 3.6 (right), where kidneys, ureters, and bladder region are represented in red, green, and blue, respectively. These maps were used in stone-synthesized augmentation in the second stage and used to evaluate the stones segmentation performance in different regions.

3.3 Synthetic Stone Augmentation

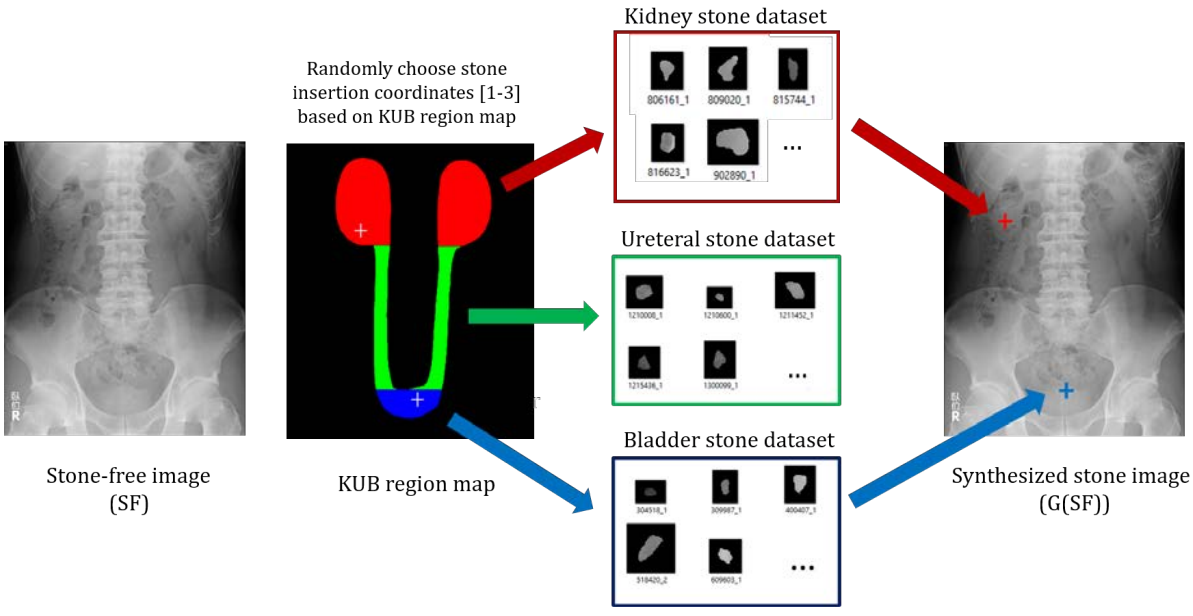


Figure 3.7: Flowchart of synthetic stone augmentation

In an abdominal x-ray image, the number of positive pixels (stone region) is very small compared to negative pixels (non-stone region). The ratio of the stone area over the non-stone area can be less than 0.1%. Furthermore, the class imbalance naturally present in many medical domains,

where "normal" images usually outnumber those with findings. Therefore, we proposed the synthetic urinary stones augmentation methods to synthetically insert the stones into healthy samples to generate new training images, as shown in the flowchart in Fig. 3.7. The proposed augmentation methods consists of urinary stone embedding algorithm and GAN-based stone inpainting augmentation. The first method is described in 3.3.2, and the latter method is described in 3.3.3.

3.3.1 Cropped stone dataset

The cropped stones were extracted from the stone-contained dataset. The KUB region map generated in the first stage were used for classifying the cropped stone into anatomic region categories consisted of kidneys, ureters, and bladder as shown in Fig 3.8 - 3.10, respectively. These cropped stone dataset will be used in stone-synthesized augmentation process.

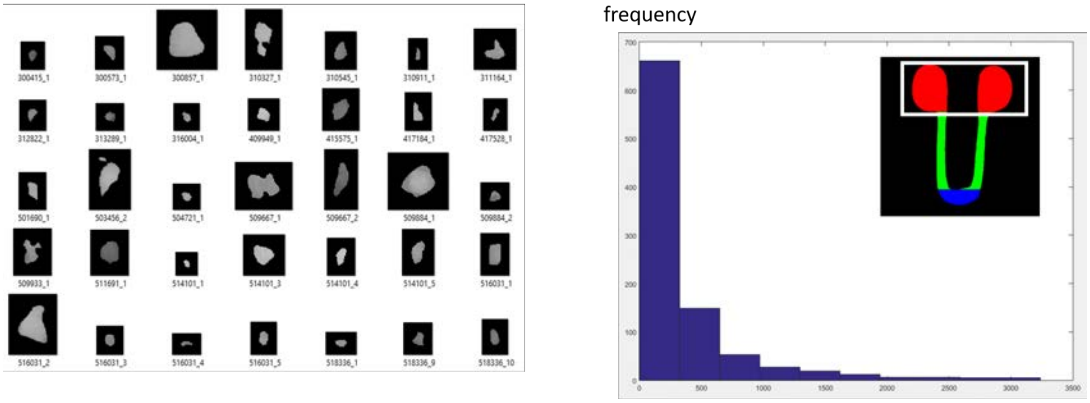


Figure 3.8: Example of cropped kidney stone dataset and the frequency distribution of their stone size.

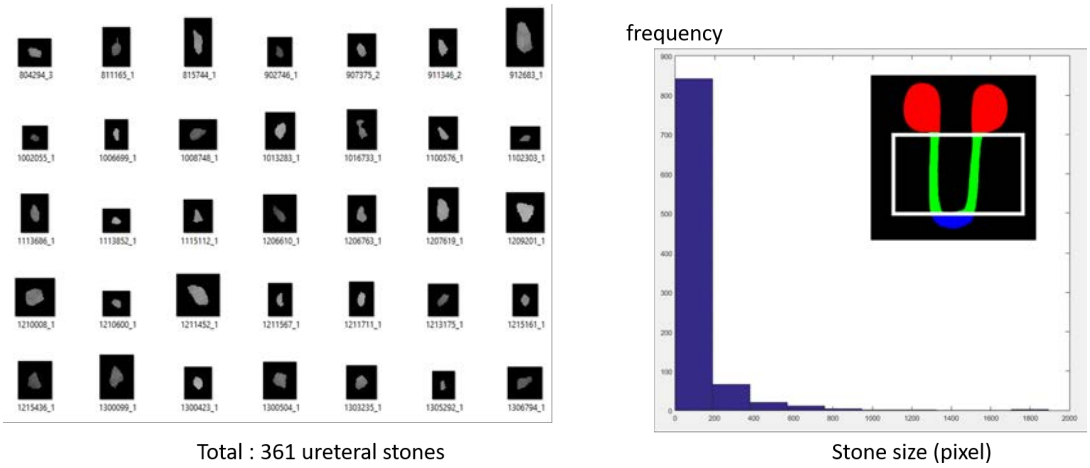


Figure 3.9: Example of cropped ureter stone dataset and the frequency distribution of their stone size.

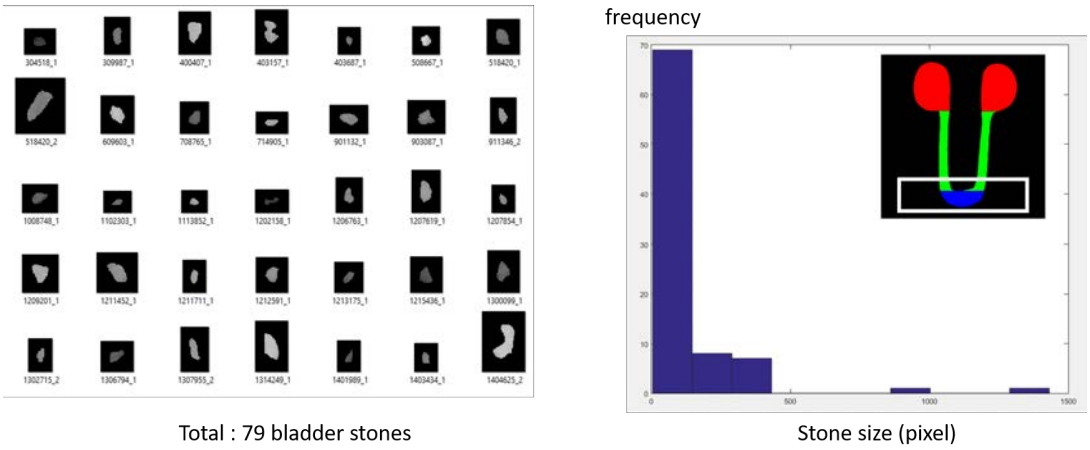


Figure 3.10: Example of cropped bladder stone dataset and the frequency distribution of their stone size.

3.3.2 Stone-embedding Augmentation

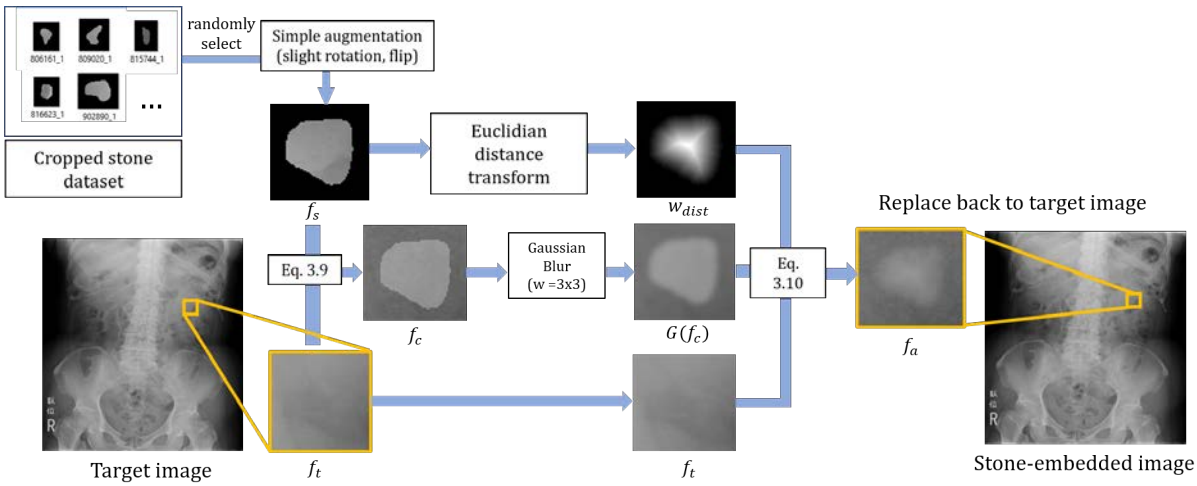


Figure 3.11: Flowchart of stone-embedding augmentation

The proposed urinary stone embedding algorithm is the method trying to generate new training images(positive) which urinary stones are inserted while preserving the background texture of the target image. Flowchart of stone-embedding augmentation is illustrated in Fig. 3.11. Firstly, urinary stone images were cropped from training stone-contained images and multiplied with cropped stone mask to remove the region outside stone pixels. Examples of cropped stones are shown in the 2^{nd} row in Fig. 3.12. Then, all stones were classified into three categories based on their location on an x-ray image: kidney stones, ureteral stones, and bladder stones. We selected the small and medium stones in a range between 20 pixels to 500 pixels which are the hard samples to use in stone-embedding augmentation.

During this augmentation process, we randomly selected 1 to 3 target location(s) (x_t,y_t) of inserted stone(s) from the KUB region map of each target image to be the center of a cropped region of the target image (f_t) that has the same size as the selected source image (f_s). Then, the source stone image f_s was randomly selected based on the region of selected locations in the KUB region map (kidneys, ureters, or bladder) and applied with simple augmentation methods

consisting of rotation [-10, 10], vertical flip, and horizontal flip. The augmented source image ($A(f_s)$) was multiplied by a stone weight λ_{stone}, which has a random value between 0.1 and 0.2, to control the brightness intensities of stone pixels and combined with f_t as shown in Eq. (3.9).

$$f_c = \lambda_{stone} A(f_s) + f_t \tag{3.9}$$

where f_c is a combined region of a source and target image.

Gaussian filter G (window size 3×3) was applied on f_c in order to making the synthetic stone looks more natural. Then, the distance map (w_{dist}) calculated by the Euclidean distance transform was used to calculate the weighted sum between $G(f_c)$ and (f_t) as shown in Eq. (3.10).

$$f_a = G(f_c) w_{dist} + f_t(1 - w_{dist}) \tag{3.10}$$

where f_a is a final stone-embedded image, which are demonstrated in Fig. 3.12 (3^{rd} row). The comparison between the embedded stones and actual stones are demonstrated in Fig.3.13. Lastly, this result is replaced back to the target image at the cropped region.

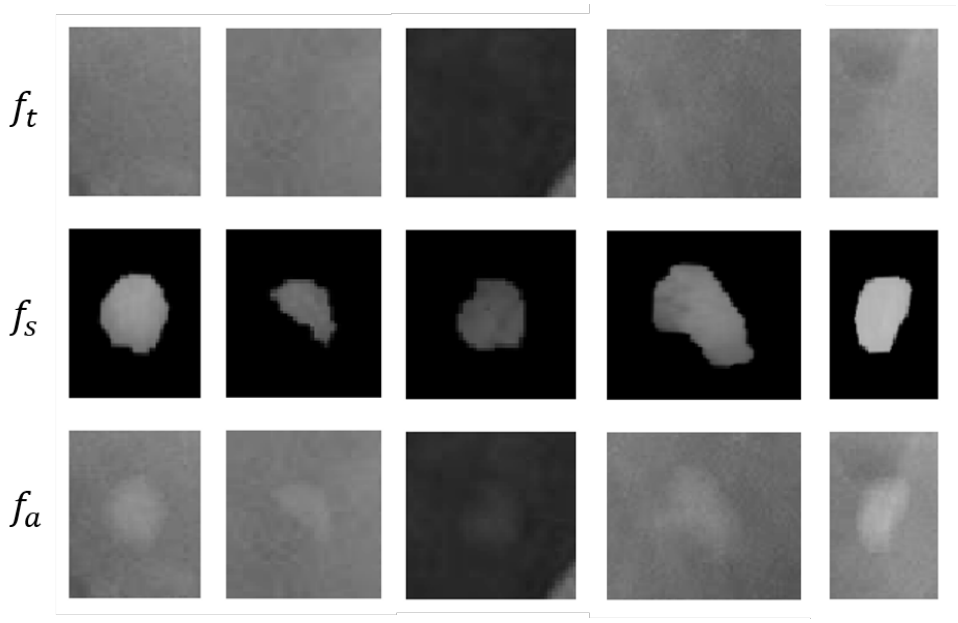
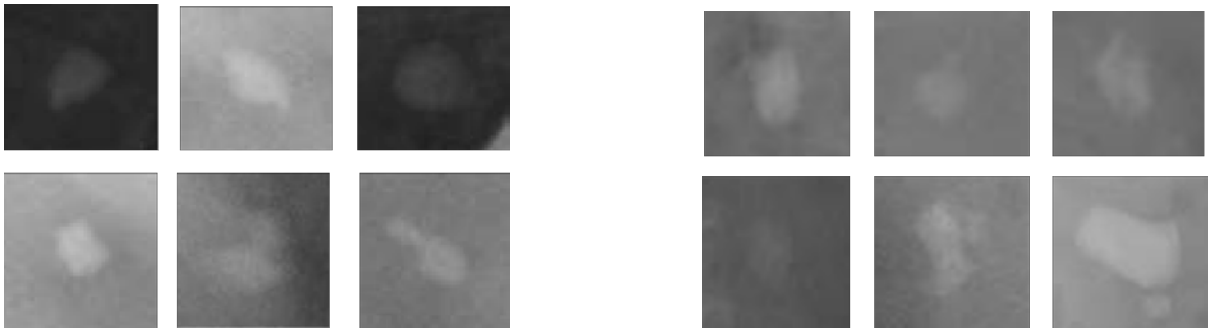


Figure 3.12: Cropped target (stone-free) images (f_t), cropped source (stone) images (f_s) and stone-embedding results(f_a).



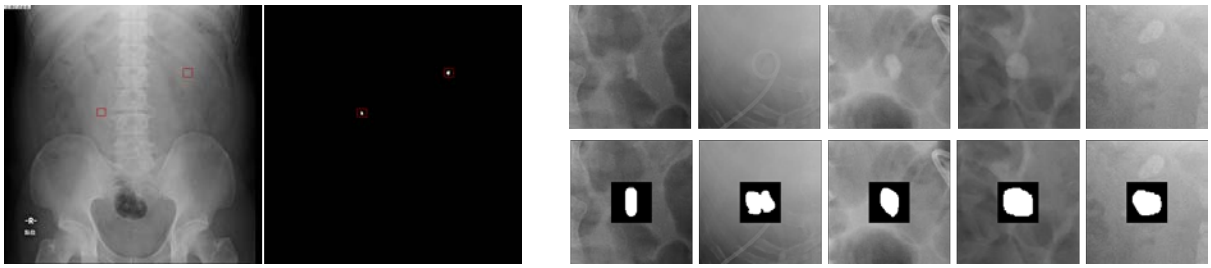
(a) Augmented Stones

(b) Real stones

Figure 3.13: The comparison between the stone-embedded images (a) and actual stone images (b).

3.3.3 GAN-based stone inpainting augmentation

Cropped stone-mask dataset



(a) Full KUB x-ray image and its ground-truth image

(b) Dataset for stone inpainting task

Figure 3.14: Illustration of an abdominal x-ray image with stones (a - left), corresponding gold standard manual segmentation of the stones (a - right). The red box represents the cropped region of urinary stone that used for creating the cropped urinary stone images and the corresponding images with binary stone mask M_s at the image's center used in stone inpainting process

In this task, abdominal x-ray images and their corresponding stone ground-truth (Fig.3.14(a)) are used to create the dataset for training image-to-image translation network. Firstly, the stone ground-truth images are cropped in square shape at the stone region for every stone, which the top-left coordinates (x_m, y_m) and the width (w_m) of stone mask M_s can be defined as Eqs.(3.11) and (3.12), respectively.

$$w_m = \begin{cases} w_s + 0.2 \cdot w_s, & \text{if } w_s \geq h_s \\ h_s + 0.2 \cdot h_s, & \text{otherwise} \end{cases} \tag{3.11}$$

where w_s and h_s are the width and height of a urinary stone region, respectively.

$$(x_m, y_m) = \begin{cases} (x_s - 0.1 \cdot w_s, y_s - \frac{w_m - h_s}{2}), & \text{if } w_s \geq h_s \\ (x_s - 0.1 \cdot h_s, y_s - \frac{w_m - w_s}{2}), & \text{otherwise} \end{cases} \tag{3.12}$$

where x_s and y_s are top-left coordinates of a urinary stone region.

Then, full abdominal x-ray images are cropped in the square shape at the stone region with the width $= 3 \cdot w_m$. Figure 3.14(b) demonstrates the original cropped stone-region images and their corresponding cropped images placing the binary stone mask at the image's center. We used these pairs for training our image-to-image translation network.

Conditional Inpainting GANs

Image inpainting is a task of reconstructing the missing or distorted region in an image. Recently, GANs have been widely used in this application instead of the traditional approaches. Context Encoder (CE) [48] is an auto-encoder architecture training with adversarial loss and reconstruction loss. The studies in [49, 50] improve the CE framework by using two discriminator networks consisting of local discriminator taking the completed region as input and global discriminator taking the entire image as input. More recently, ip-MedGAN [51] is the inpainting

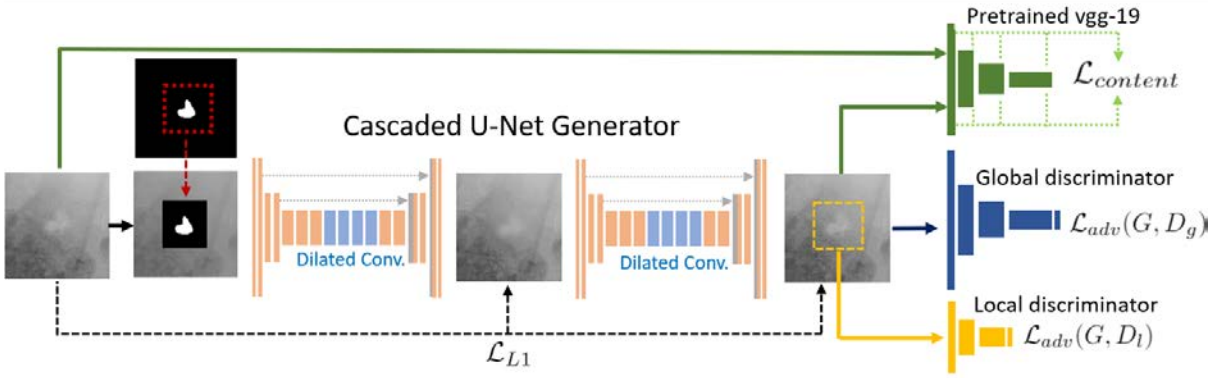


Figure 3.15: An overview of our generative stone inpainting framework. The cascaded U-Net generator using dilated convolution is trained with reconstruction loss, content loss from pre-trained VGG19, global adversarial loss, and local adversarial loss.

framework developing for medical imaging. This method uses cascaded multiple U-Net networks as the generator trained with the combination loss of discriminator networks, reconstruction loss, perception loss, and style loss.

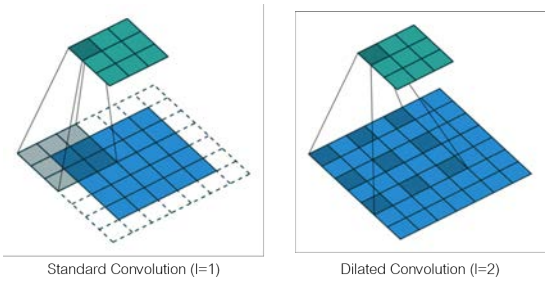
In this work, we used the similar image-to-image translation network to generated the missing region at the stone mask input. This conditional GAN, learning a mapping from observed image x and random noise z to y, has two components including a generator and a discriminator. The generator G is trained to generate the output images, which are difficult to be distinguished from real images, while the discriminator D is trained to classify between the fake generated images and real images. The adversarial loss of a conditional GAN can be expressed as

$$\mathcal{L}_{cGAN} = E_{x,y}[logD(x,y)] + E_{x,y}[log(1 - D(x, G(x,z)))] \tag{3.13}$$

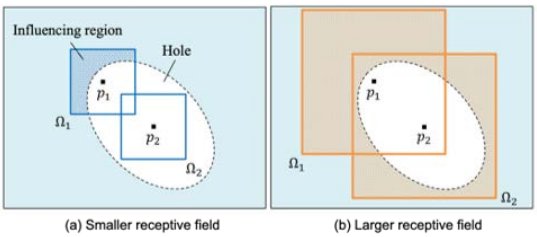
where G tries to minimize this objective, while an adversarial D tries to maximize it.

Generator architecture

The overall framework of this image-to-image translation network is illustrated in Fig.3.15 We used the stack of two U-Net models, as the inpainting generator, which the input for second network is the coarse inpainted result of the first network. Each model has two paths consisting of contracting path and expanding path. The generator receives 128 $\times$ 128 full images with masked region as the input. Each convolutional block consists of two 3 $\times$ 3 convolutional layers with LeakyReLu activation and Batch normalization, followed by a 3 $\times$ 3 convolutional layers with stride of 2 to downsample the image resolution. At the mid-layers, we use the dilated convolutional layers with dilation rate (η) of 2, 4, 8, and 16. Dilated convolution increases the receptive field, while still using the same number of parameters and computational resources. These layers at the low resolution is important for image inpainting task because it needs a larger receptive field that can cover the contextual information and missing region as shown in Fig.3.16. In the expanding path, transposed convolutional layer is implemented to upsample the image resolution and concatenated with the encoder at the same spatial level. The output layer of each generator uses a 1 $\times$ 1 convolutional layer with tanh activation.



(a)



(b)

Figure 3.16: The illustration of standard convolution and dilated convolution (a) and the importance of dilated convolution in image inpainting task (b).

Discriminator architecture

Image inpainting task usually utilizes two discriminators with different receptive fields. The global discriminator D_g receives the full generated and real images as the input, like other GANs, while the local discriminator D_l receives only the masked region of generated and real images as the input. The global discriminator network has the receptive field of 128 $\times$ 128 pixels and consists of 4 convolutional layers(convolution + LeakyReLU + Batch normalization) with 2 strides. By using the wide receptive input, the network focuses on the realistic in entire image detail and ensure that the inpainted region fits to the contextual information around the masked region. The local discriminator network has the receptive field of 48 $\times$ 48 pixels cropped at the masked region and consists of 3 convolutional blocks (convolution + LeakyReLU + Batch normalization) with 2 strides. By using the smaller input receptive field at the masked region, this network only focuses on the realistic within the inpainted region. The last layer of both network is a 1 $\times$ 1 convolutional layer with Sigmoid activation, which produce the output patch N $\times$ N representing the classification scores ('real' or 'fake'). The adversarial loss of cGAN ($\mathcal{L}_{adv}$) used in this work is the average between these two discriminators with different receptive fields, which can be expressed as

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}(G, D_g) + \mathcal{L}_{adv}(G, D_l) \tag{3.14}$$

Training Methodology

Recently, non-adversarial losses were usually used in image-to-image translation task as it can obtain the better and consistent results [37]. In this work, we used a conventional pixel-wise reconstruction loss ($\mathcal{L}_{L1}$) as shown in Eq. (3.15) to minimize the mean absolute error (MAE) between the target and generated image.

$$\mathcal{L}_{L1} = E_{x,y,z}[\|x - G(y, z)\|_1] \tag{3.15}$$

We also utilized the content loss to enhance image details of inpainted image. The feature maps of target $V_j(x)$ and generated image $V_j(y, z)$ were extracted from different j^{th} convolutional layers of a pre-trained VGG-19 network trained on the ImageNet dataset in the classification task []. Then, $\mathcal{L}_{content}$ can be computed by

$$\mathcal{L}_{content} = \sum_{j=1}^{B} \frac{1}{h_j w_j d_j} \|V_j(x) - V_j(G(y, z))\|_2 \tag{3.16}$$

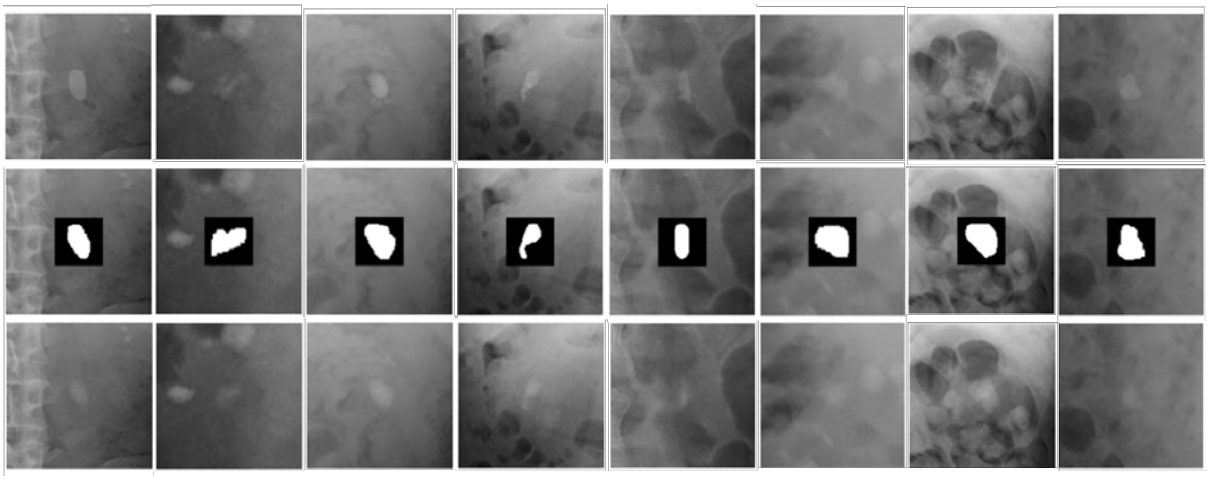


Figure 3.17: The illustration of original cropped stone region images (1^{st} row images), input images for stone inpainting network(2^{nd} row images), and synthesized urinary stone results generated by stone inpainting network (3^{rd} row images).

where the feature maps have the size $h_j \times w_j \times d_j$ with h_j,w_j, and d_j being the height, width, and depth, respectively.

The first part of the cascaded U-Net model was trained with only $\mathcal{L}_{L1}$ loss to generate the coarse result, while the second network was optimized by using the combined objective functions of adversarial loss, L1 reconstruction loss, and content loss expressed as:

$$\lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_1 + \lambda_3 \mathcal{L}_{content} \tag{3.17}$$

where $\lambda 1$, $\lambda 2$, and $\lambda 3$ represent the contributions of adversarial loss, L1 loss, and content loss, respectively.

We used ADAM optimizer [64] with momentum value of 0.5 and a learning rate of 0.0002 to train the network for 15,000 iterations. The discriminator was trained once for every two iterations of training the generator. The examples of synthesized urinary stone result from stone inpainting network are illustrated in Fig. 3.18, which the pixels in the missing region is filled based on the binary stone mask input.

3.3.4 Stone-synthesized Dataset

In an abdominal x-ray image, the number of the positive pixels (stone region) is very small compared with the negative pixels (non-stone region). The ratio of the stone region over the non-stone area can be less than 0.1%. In this stage, we use the proposed urinary stone inpainting method described in the previous section to increase the number of positive data. The proposed framework of GAN-based synthetic stone augmentation is illustrated in Fig.3.18. We implemented this proposed augmentation to both stone-free image (I_{sf}) and stone-contained image (I_{sc}).

For each real stone-free image (I_{sf}), 1 to 3 new target location(s) (x_t, y_t) were randomly selected from the corresponding KUB region map to synthesize new stone(s) in the non-stone region. A cropped stone mask (M_s) was randomly selected from the cropped stone mask dataset and augmenting using image rotation [-10 , +10], vertical flipping, and horizontal flipping to increase the diversity of the new stone's characteristics. The augmented stone mask M_s was then placed in the center of a selected location (x_t, y_t), and a full image (I_{sf}/I_{sc}) was cropped

in a square shape around the placed stone mask M_s with a $3 \cdot w_m$ width to include the context region surrounding the stone mask, similar to the training data for the stone inpainting task. For each real stone-contained image (I_{sc}), the center coordinate of each stone (x_t, y_t) was randomly chosen to be replaced with either the stone mask to synthesize a new stone, or non-stone mask to remove the stone when there are multiple stones.

The input is resized to 128 × 128 pixels and processed by the inpainting generator that was trained in the first stage to generate the stone region based on the context pixels around the missing region and an input stone mask. The example of stone-synthesized results of I_{sc} and I_{sf} are illustrated in Fig. 3.19 (a) and (b), respectively. Then, the inpainted result is replaced at the cropped region in the full image, and the stone mask is placed at the same location in the full ground-truth image. This augmentation method was implemented to generate 10 new samples for each I_{sc} and I_{sf}, which was used as the additional training samples for segmentation network.

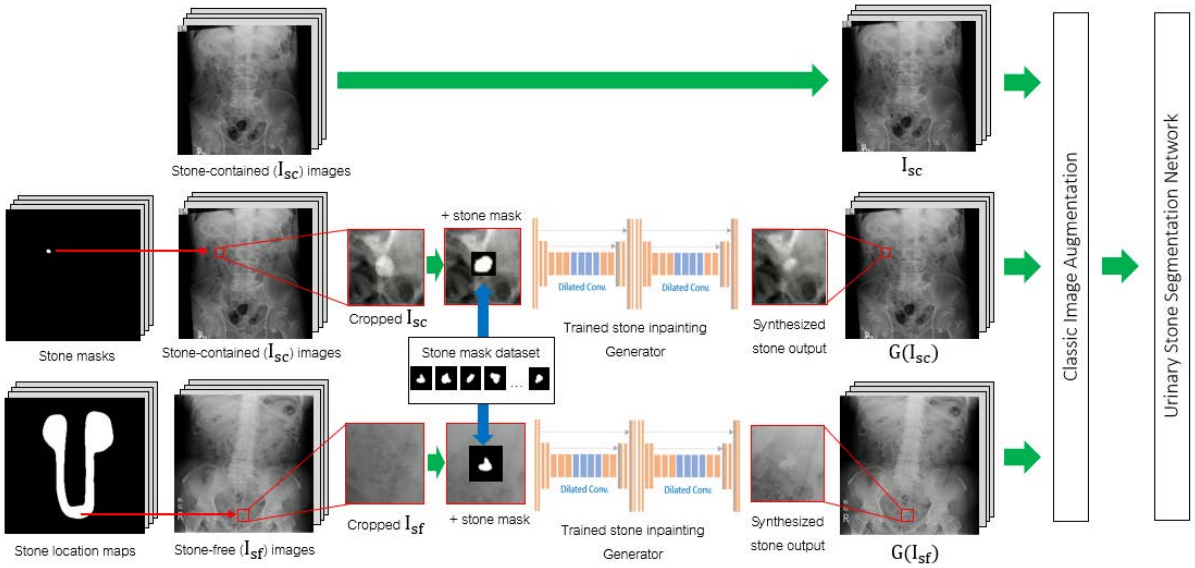
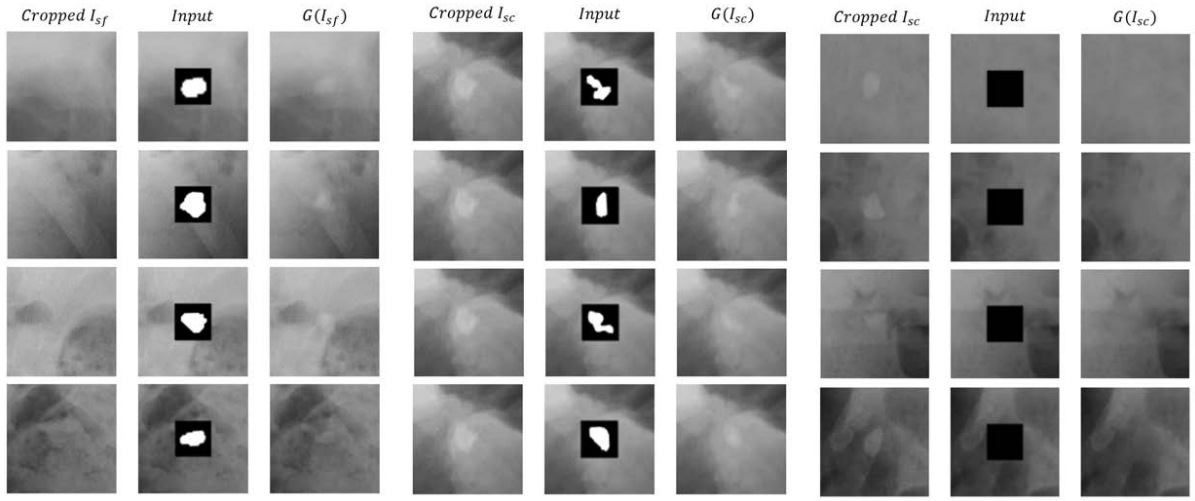


Figure 3.18: The proposed framework of image augmentation consisted of GAN-based augmentation and classic augmentation for urinary stone segmentation.



(a)

Figure 3.19: Illustration of original cropped target images, cropped targetimages with random masks, and stone-inpainted results in stone-free images(a), and stone-contained images (b).

3.4 Urinary Stones Segmentation

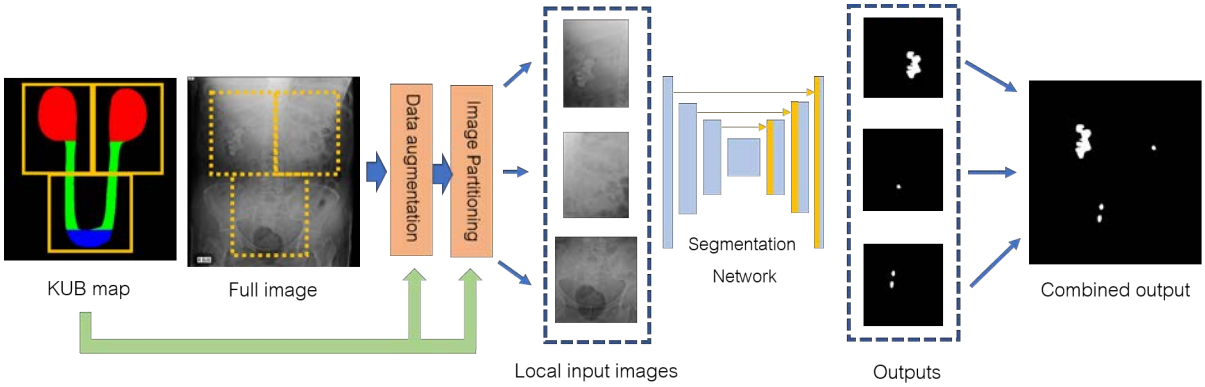


Figure 3.20: Flowchart of urinary stone segmentation. The segmentation network receives the partitioned input. The partitioned output are combined into full image results in the final step.

3.4.1 Pre-processing and Image Partitioning

The luminance and contrast of our x-ray images are varied because of different conditions during the image acquisition process. Therefore, all samples were normalized to a zero mean and unit variance before the training process, While all ground-truth images were converted to binary images where 1s pixels represent the stone region, and 0s pixels represent the background region. Then, all full images were partitioned based on the stone location component into 3 local images including top-left partition (P_{tl}), right partition (P_{tr}), and bottom partition (P_b) by Eqs.(2) - (5) as described in the first-stage section. Finally, all training and ground-truth partitioned images were resized to 256×256 before training.

3.4.2 Reweighting approach to balance stone size inequality

Lesion size inequality problem

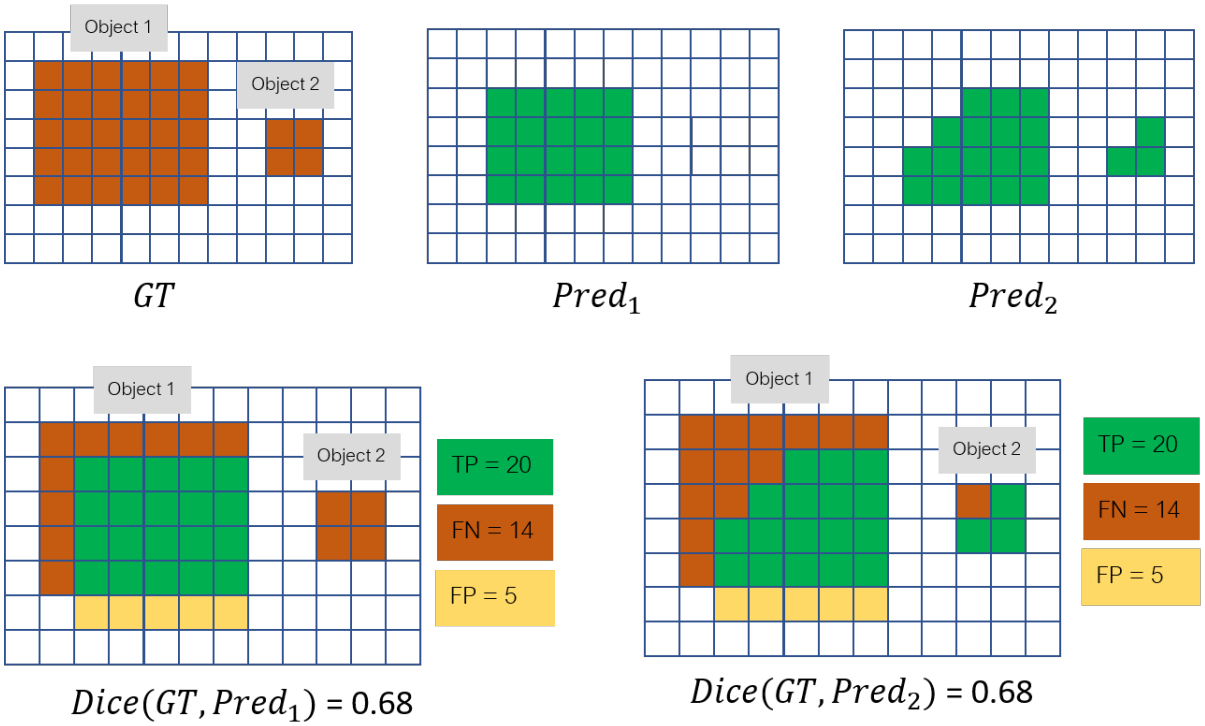
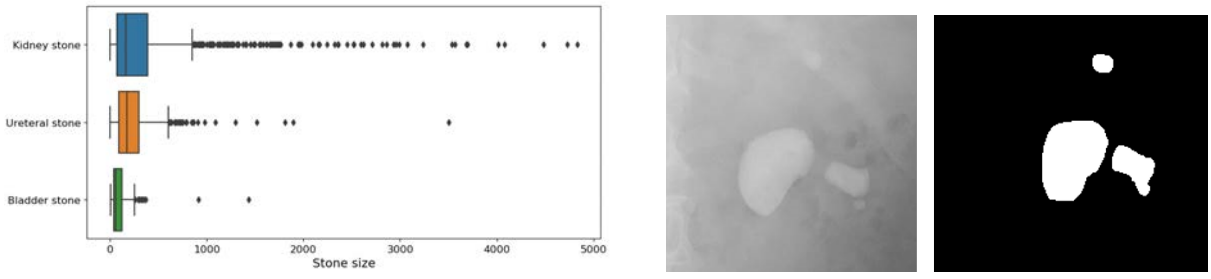


Figure 3.21: The illustration explaining the lesion size inequality problem when using simple pixel-wise segmentation metric.

Assume that the ground-truth GT has two objects and we have two predicted results as show in Fig. 3.21. If we compute the dice coefficient score between the GT and two prediction examples, two examples has the same segmentation score. However, only large object can be detected in example 1, while two objects in example 2 are detected. Problem of pixel-wise segmentation score such as dice coefficient is that big lesion overshadows small one.



(a) (b)

Figure 3.22: Stone sizes distribution (a), and an example of cropped stones region showing stone size imbalance.

From our preliminary experiments, the model trended to miss small stones when training with the traditional dice coefficient loss or binary cross entropy. The main reason is that large

lesions overshadow the small ones in loss calculation. Most of the recently proposed loss functions try to solve the data imbalance problem between classes [611] by adding weight to a loss function to balance each class, but ignore imbalance between lesion size in the same class. In our case, abdominal x-ray images usually have multiple stones per image, and some stones can be more than 100 times larger than small ones, as shown in stone sizes distribution in Fig.3.22(a).

Proposed lesion reweighting approach

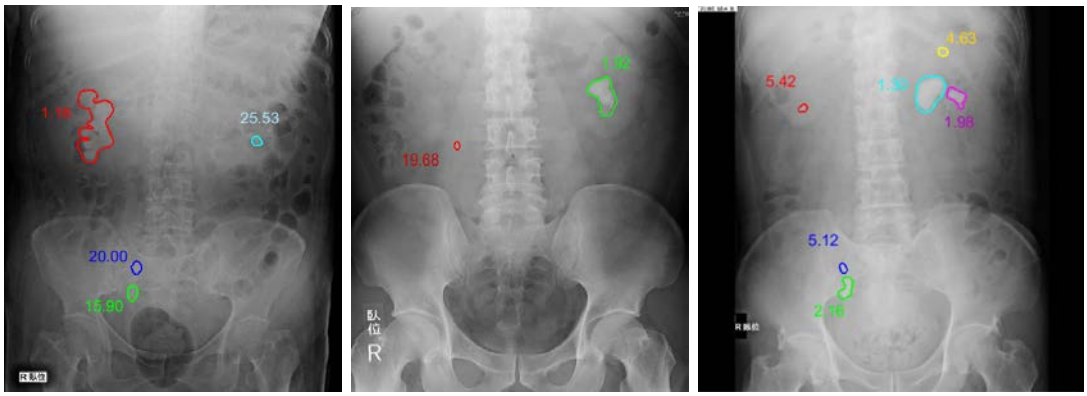


Figure 3.23: Illustration of an inverse weighting result calculated using our lesion reweighting method. A Weight for every stone is shown near the stone contour.

In this work, we proposed the stone size reweighting method, which is inspired from [62], to reduce the lesion size imbalance problem during training process. The difference is that our inverse weighting method does not include the background component because highly imbalance between the background and stone region makes the weights of stone pixels too high, which reduced segmentation performance in our case. During the training process, we generated the tensor of weights for every batch. We split a tensor of ground-truth into N 2-D connected components. Then, the weight for every pixel inside each component (w_j) can be computed by Eq.(3.18).

$$w_j = \begin{cases} 1, & \text{if } j = 0 \\ 1 + \dfrac{\sum_{n=1}^{N} |C_n|}{N \cdot |C_j|}, & \text{otherwise} \end{cases} \tag{3.18}$$

where C_0 is the background component, and $C_1, \ldots, C_N$ are the connected components [65] of stones in the current batch. This inverse weighting method assigns the higher weights to small stones (Fig.3.23), which will be used in loss calculation during the training stage.

3.4.3 Training methodology

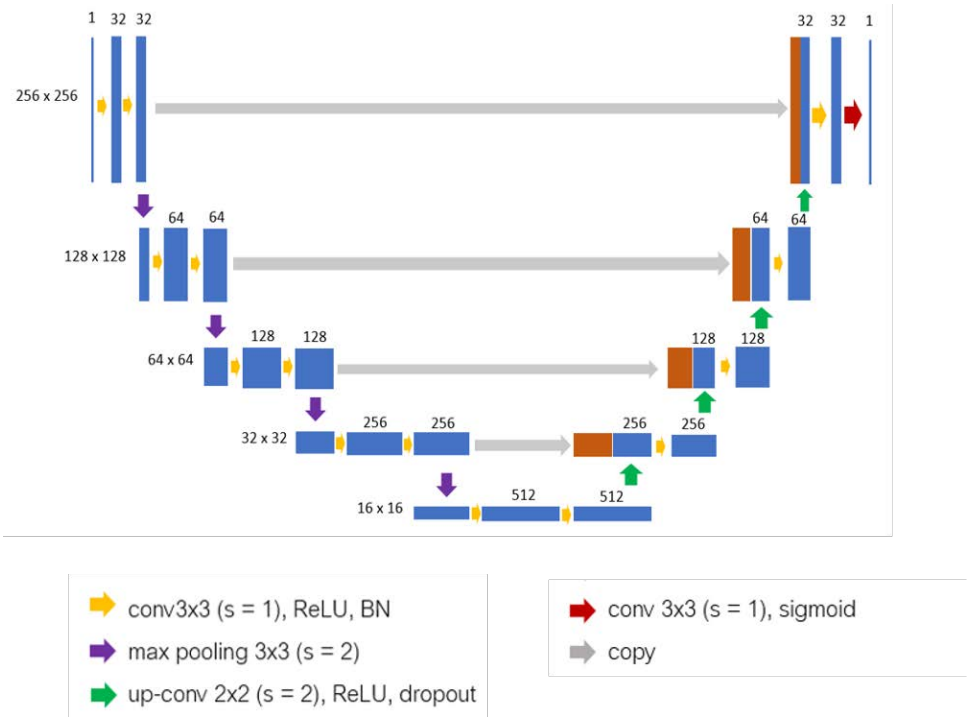


Figure 3.24: Our U-Net architecture.

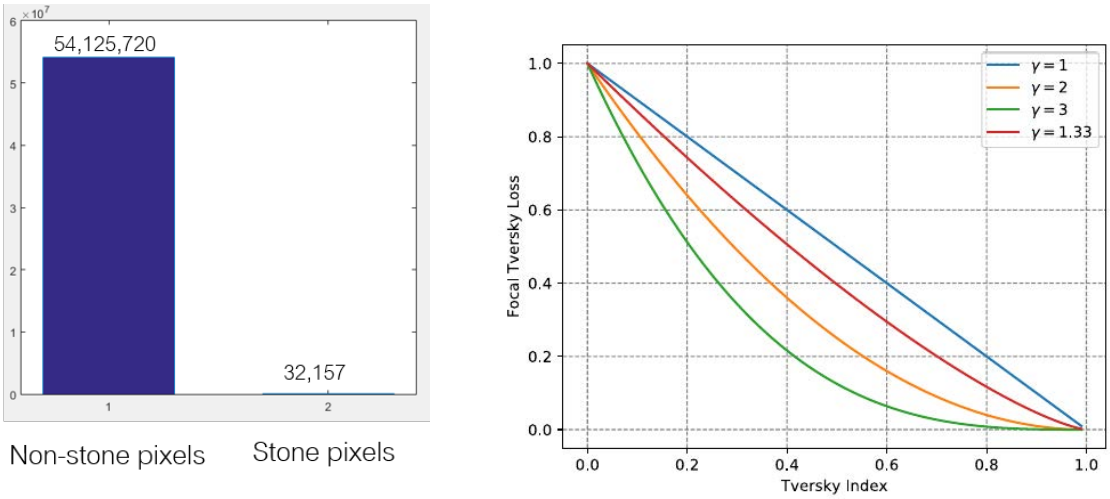
In this stage, training images consist of all stone-contained images (I_{sc}) images, stone-synthesized stone-contained images ($G(I_{sc})$), and stone-synthesized stone-free images ($G(I_{sf})$). Only 1/4 of the $G(I_{sc})$ and $G(I_{sf})$ training samples were randomly selected in each epoch and combined with the real training samples I_{sc}. The training images were implemented basic augmentation methods including image rotation [-5, 5], and horizontal flipping. We selected these augmentation methods because they are not distorted the image information, and they are also suitable for our dataset. Then, all training samples were normalized and cropped into 3 partitions using the corresponding KUB region maps as described in the pre-processing section.

The same U-Net model in the first stage was also used in this task. The architecture of U-Net model that we used is shown in Fig. 3.24, which the detail is defined in Table 3.1. The input image (256×256×1) were normalized to [-1, 1] before training, and the output image (256×256×1) from the Sigmoid activation function is the continuous values between 0 and 1 representing the probability of urinary stone pixels.

Table 3.1: The detail architecture of the U-net model for urinary stone segmentation.

Layers name	Type	Output shape
Input	Image	256 x 256 x 1
conv2d_1	2D Convolution (3x3)	256 x 256 x 32
conv2d_2	2D Convolution (3x3)	256 x 256 x 32
max_pooling2d_1	Max Pooling (3x3, s=2)	128 x 128 x 32
conv2d_3	2D Convolution (3x3)	128 x 128 x 64
conv2d_4	2D Convolution (3x3)	128 x 128 x 64
max_pooling2d_2	Max Pooling (3x3, s=2)	64 x 64 x 64
conv2d_5	2D Convolution (3x3)	64 x 64 x 128
conv2d_6	2D Convolution (3x3)	64 x 64 x 128
max_pooling2d_3	Max Pooling (3x3, s=2)	32 x 32 x 128
conv2d_7	2D Convolution (3x3)	32 x 32 x 256
conv2d_8	2D Convolution (3x3)	32 x 32 x 256
max_pooling2d_4	Max Pooling (3x3, s=2)	16 x 16 x 256
conv2d_9	2D Convolution (3x3)	16 x 16 x 512
conv2d_10	2D Convolution (3x3)	16 x 16 x 512
conv2d_transpose_1	2D Convolution Transpose (2x2, s=2)	32 x 32 x 256
concateneted_1	Concatenete	32 x 32 x 512
conv2d_11	2D Convolution (3x3)	32 x 32 x 256
conv2d_transpose_2	2D Convolution Transpose (2x2, s=2)	64 x 64 x 128
concateneted_2	Concatenete	64 x 64 x 256
conv2d_12	2D Convolution (3x3)	64 x 64 x 128
conv2d_transpose_3	2D Convolution Transpose (2x2, s=2)	128 x 128 x 64
concateneted_3	Concatenete	128 x 128 x 128
conv2d_13	2D Convolution (3x3)	128 x 128 x 64
conv2d_transpose_4	2D Convolution Transpose (2x2, s=2)	256 x 256 x 32
concateneted_4	Concatenete	256 x 256 x 64
conv2d_14	2D Convolution (3x3)	256 x 256 x 32
conv2d_15	2D Convolution (3x3)	256 x 256 x 32
conv2d_16	2D Convolution (3x3) + Sigmoid	256 x 256 x 1

The loss function is very important for training deep learning models. In the research related to medical images, the Dice score coefficient (DSC), an overlap index, has been widely used to optimize the segmentation performance. However, the limitation of the Dice loss function is that it weights false negative (FN) and false positive (FP) equally, which results usually have high precision but low recall [59]. Moreover, it struggles to segment small ROIs (urinary stones) because they are very small compared with the image background (Fig. 3.25(a)) and do not significantly contribute to the network's loss. Another challenge is that the stones usually have various sizes which the large stones overshadow small stones in the training process, which is described in the previous section.



(a) (b)

Figure 3.25: Number of non-stone pixels and stone pixel (a), and the plot between focal Tversky loss and Tversky index.

In this work, we used focal Tversky loss (FTL) applied with inverse weighting map (iw) to overcome these challenges. Focal Tversky loss is the generalization of the Dice loss (DL) balancing importance between FN and FP by α and β, respectively. Furthermore, it also has γ for controlling between easy and hard training samples [60]. When $\gamma > 1$, this loss focuses more on small TI samples that are misclassified (Fig.3.25(b)). The calculation of TI_{iw} and FTL_{iw} is defined as Eq. (3.19) and Eq. (3.20), respectively.

$$TI_{iw} = \frac{\sum_{i=1}^{N} w_i p_{1i} g_{1i} + \epsilon}{\sum_{i=1}^{N} w_i p_{1i} g_{1i} + \alpha \sum_{i=1}^{N} w_i p_{0i} g_{1i} + \beta \sum_{i=1}^{N} w_i p_{1i} g_{0i} + \epsilon}, \tag{3.19}$$

$$FTL_{iw} = (1 - TI_{iw})^{1/\gamma} \tag{3.20}$$

where p_{1i} is the probability of pixel i being a stone and p_{0i} is the probability of pixel i being a non-stone. While g_{1i} is 1 for a stone pixel and 0 for a non-stone pixel and g_{0i} vice versa. Total number of pixels in a current batch is denoted by N. ϵ is added in numerator and denominator term in both formula to avoid division by zero.

This work assumed that higher α can improve the performance by reducing FN; therefore, $\alpha = 0.7$ and $\beta = 0.3$ were used for weighting more on FN than FP. After we experimented with high values of γ, $\gamma = 2.0$ had the best performance in our case; thus, we used these values in all experiments.

We trained the network from scratch for every experiment and used Adam optimizer to minimize FTL_{iw} with an initial learning rate of 10^{-3}. During training, whenever validation loss has not decreased by at least 10^{-4} for 10 epochs, the learning rate is divided by 2 while the minimum learning rate is set as 5×10^{-4}. Model was trained for 150 epochs with a batch size of 12 images for all experiments.

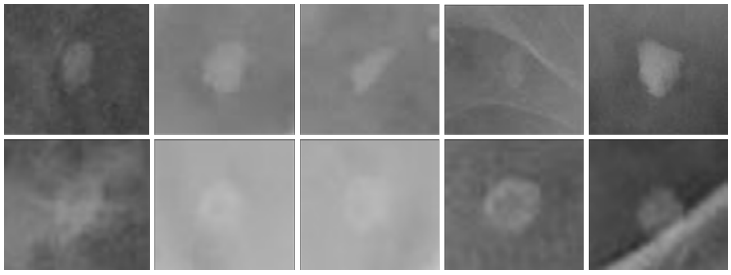


Figure 3.26: Illustration of a comparison between bladder stones (1^{st} row) and phleboliths (2^{nd} row) from our dataset.

3.4.4 Post-processing Stage for False Positive Reduction

Phleboliths

Calcifications of tiny veins or phleboliths, as shown in Fig. 3.26 (bottom), are prevalent in bladder region and can be difficult even for an expert to identify from urinary stones in this location (top). Several studies have reported that urinary stones and phleboliths present different morphological structures and characteristics, however, the classification is still challenging especially for the x-ray modality [66, 67].

Based on our preliminary experiments, the false-positive results usually occurred in the bladder region, which has two main reasons: the similar characteristics of phleboliths and the limited positive samples in this region. To solve these problems, our work proposed GAN-based stone inpainting augmentation to increase the amount of samples in this region, and bladder stones classification to reduce the false-predictions, which will be described in the following sections.

Pretrained Model for Bladder stones detection

We proposed the detection of false bladder stone by training the classification model to distinguish between bladder stones and phleboliths. To prepare the dataset using for training the network, We manually cropped 150 images of the bladder stone and phleboliths as well as the paired stone masks from the training images dataset.

The pre-trained VGG16 network (Fig.3.27), which were trained by ImageNet dataset, was used for training and fine-tuned with our dataset only for the fully-connected (FC) layers. The input image was the concatenation of the cropped image, stone mask input, and zeros images, and was resized to 224×224 pixels. The pre-trained model was trained using Focal binary cross entropy loss for 150 epochs. Image augmentation methods including image rotating [-20, 20], horizontal flipping, vertical flipping, and image zooming [0, 0.1] were implemented during the training process to increase the number and diversity of training images.

Post-processing

In post-processing stage, we used the trained bladder stone classification model to remove the false-positive predictions in the bladder partition from the segmentation map of the 2^{nd} stage segmentation network. The candidate lesions (bladder stones) were cropped by the each connected components, and used as the inputs for bladder stones detection network. The candidate lesions that has the predicted score lower than 0.5, which predicted as the false prediction, will

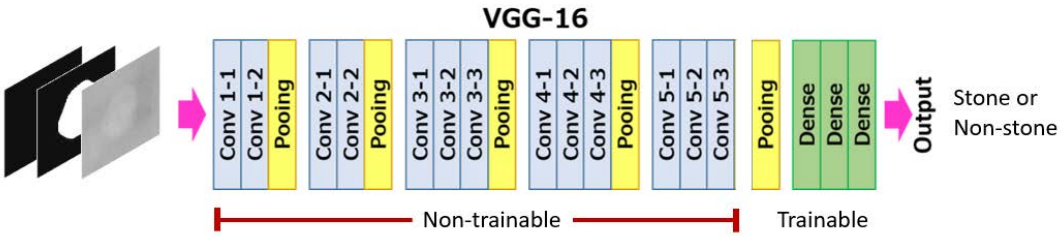


Figure 3.27: Bladder stone detection using pretrained VGG16 model.

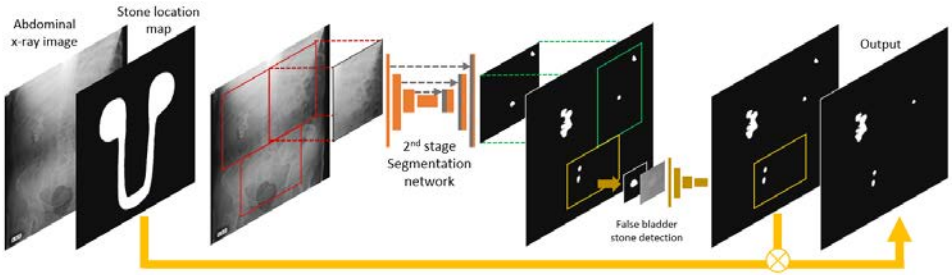


Figure 3.28: Bladder stone detection using pretrained VGG16 model.

be removed. Lastly, the full-scale segmentation outputs then multiplied with the corresponding stone location maps produced from the 1^{st} stage segmentation network to remove the false predicted lesions outside of the urinary organ region.

Chapter 4

Evaluation

4.1 Evaluation methods

4.1.1 Pixel-wise Evaluation

As our dataset is a high imbalance problem, a simple pixel-wise accuracy metric, comparing every pixel between the prediction and gold standard stone mask is not suitable in our case (the blackground pixel will dominate the overall result). In common image segmentation task, the segmentation results are evaluated by segmentation-based results, which focus only the foreground object(s), and ignore the background component. Therefore, we can obtain these confusion matrix including:

1.) True positive (TP) A test pixel that correctly predicts as the stone pixel.

2.) False negative (FN) A test result which wrongly predicts as the non-stone pixel, but the ground-truth is stone pixel.

3.) False positive (FP) A test result which wrongly predicts as the stone pixel, but the ground-truth is non-stone pixel.

Then, the pixel-wise metrics include recall, precision, and F-score can be computed using TP, FN, and FP values as indicated in Eqs.(4.1 - 4.3).

$$Recall = \frac{TP}{TP + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$F_B = \frac{(1 + B^2) \cdot Precision \cdot Recall}{(B^2 \cdot Precision) + Recall} \tag{4.3}$$

Because the stone pixels are more important than the non-stone pixels, F_2 is suitable in our case.

4.1.2 Region-wise Evaluation

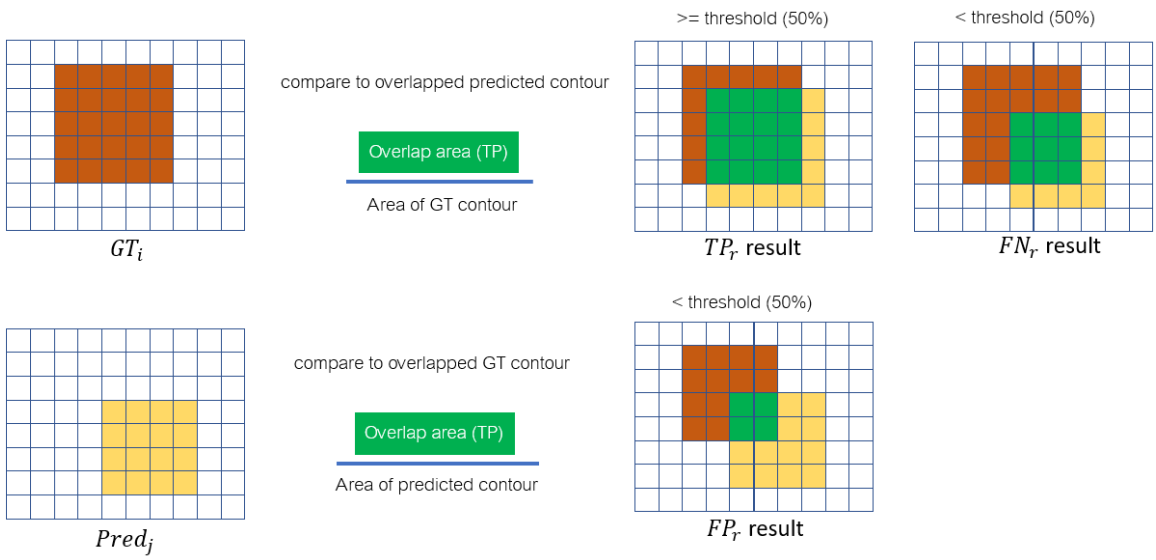


Figure 4.1: Region-wise evaluation method.

Although conventional pixel-wise evaluation has been used in many segmentation tasks, it has a drawback in the multiple lesion detection because big lesions overshadows the small ones. Therefore, we also evaluated the results by using the region-wise metrics, measuring the detection performance based on the ground-truth stones and predicted stones.

In every testing image, each connected component of stone-ground truth (G_i) is compared with the predicted stone connected component P that overlaps G_i. Total region-wise true positive (TP_r), false negative (FN_r) can be defined in Eqs. 4.4, and 4.5, respecively.

$$TP_r = \sum_{i=1}^{N} G_i[\frac{G_i \cap P}{G_i} >= 0.5] \tag{4.4}$$

$$FN_r = \sum_{i=1}^{N} G_i[\frac{G_i \cap P}{G_i} < 0.5] \tag{4.5}$$

where the stone ground-truth have total N connected components .

To compute false positive (FP_r), each predicted connected component (P_j) is compared with the ground truth that overlaps P_j. Then, FP_r can be defined as Eq. 4.6.

$$FP_r = \sum_{j=1}^{M} P_j[\frac{P_j \cap G}{P_j} < 0.5] \tag{4.6}$$

where the predicted stones have total M connected components.

By using the region-wise evaluation metric, the difference between lesion's size does not effect to these scores. Then, TP_r, FN_r, and FP_r were used to compute region-wise recall, precision, and F score, as display in Eqs. (4.7 - 4.9), respectively. We selected to use F_2 score that has a higher weight in FN_r than FP_r; as we focused more on the detection of urinary stones, which some increasing false positive as a trade-off was acceptable.

$$Recall = \frac{TP_r}{TP_r + FN_r} \tag{4.7}$$

$$Precision = \frac{TP_r}{TP_r + FP_r} \tag{4.8}$$

$$F_B = \frac{(B^2 + 1) \cdot Precision_r \cdot Recall_r}{(B^2 \cdot Precision_r) + Recall_r} \tag{4.9}$$

4.1.3 T-Test : Comparing Group Means

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups.

1.) The null hypothesis (H0) : "the two population means are equal" ($\mu_1 = \mu_2$)

2.) The alternative hypothesis (H1) : "the two population means are not equal" ($\mu_1 \neq \mu_2$)

where μ_1 and μ_2 are the population means for group 1 and group 2, respectively.

This statistical analysis is useful for comparing the two groups results from cross validation experiments. In this work, we used independent samples t-test which is defined as Eq. (4.10).

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{4.10}$$

where

$\bar{x}$ = Mean of first sample

$\bar{y}$ = Mean of second sample

n = Sample size of first sample

m = Sample size of second sample

s_x = Standard deviation of first sample

s_y = Standard deviation of second sample

The calculated t value is compared with the critical t value from the t-distribution (Fig.4.2) with degrees of freedom (n + m - 2), and confidence level ($p < 0.05$). If the calculated t value $>$ critical t value, then we reject the null hypothesis, meaning that the two population means are not equal.

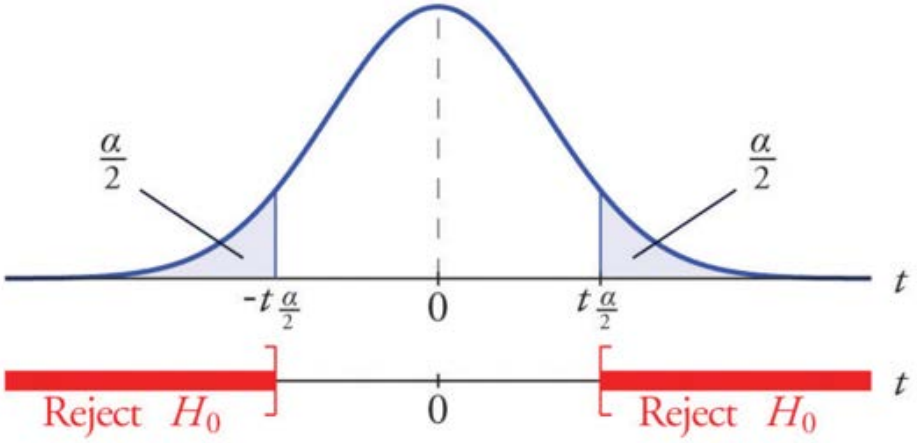


Figure 4.2: T-distribution

4.2 Stone Location Map Experiments

4.2.1 Experiment setup

We used total 750 full plain x-ray images in the first stage experiment; consisted of 70% training images, 10% validating images, and 20% testing images. All testing images were processed by the 1^{st} stage U-Net model and converted to the binary images using a 0.5 threshold value. Then, we evaluated the results of this task using simple pixel-wise metrics including recall, precision, and F_1 score.

4.2.2 Results and Discussion

Training and validating dice coefficient loss graph of stone location map segmentation is displayed in Fig.4.3. The training and validating loss was decreased rapidly in the first 20 epochs. Then, the validating loss was almost stopped decreasing, while training loss was continue decreasing (over-fitting).

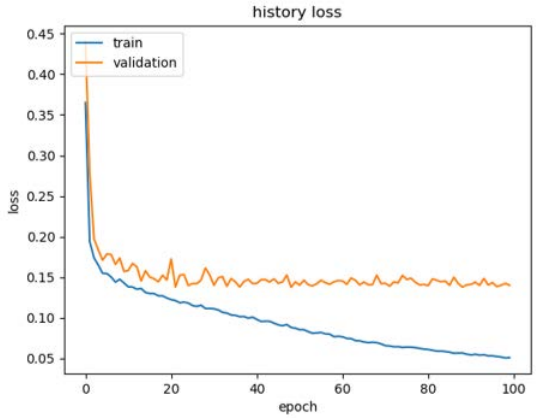


Figure 4.3: Training and validating dice coefficient loss graph of stone location map segmentation.

Table 4.1: Pixel-wise evaluation of the stone location map segmentation measured by recall, precision, and F_1 score (average ± S.D. %).

	Recall (%)	**Precision (%)**	**F_1 score (%)**
	Average (±S.D.)	**Average (±S.D.)**	**Average (±S.D.)**
1st stage U-Net	0.84 (±0.05)	0.90 (±0.04)	0.87 (±0.03)

Pixel-wise the stone location map segmentation results (mean ± s.d.) is presented in Tables I. Our first stage U-Net can produce 0.84% recall, 0.90% precision, and 0.87% F_1 score. Examples of stone location map result are displayed in Fig.4.4. The top-five best results, showing in the first row, demonstrate that these maps can represent the kidneys, ureters, and bladder region and our stone location map generated by the U-Net model corresponds to the characteristics of input abdominal x-ray images. The top-five lowest F-score results, showing in the second row, demonstrate that although the stone location map results are not segmented precisely compared with the ground-truth, the overall results can still represent the approximate location of the urinary organs and have the dice coefficient more than 0.79 for every testing images, which is acceptable result in this stage. Example of plain abdominal x-ray images (top), and their KUB region maps (bottom) are shown in Fig.4.5. From these example results, the shape of KUB region map created by the generated stone location map are varied based on the characteristics of input abdominal x-ray images.

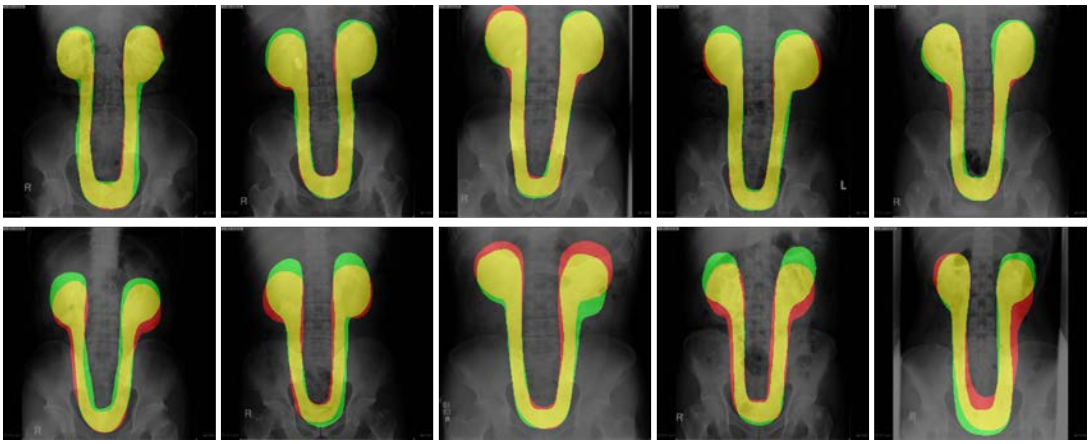


Figure 4.4: Illustration of stone location map results from the 1^{st} stage U-NET; plain x-ray images are overlaid with the predicted map and ground-truth map where TP, FP, and FN pixels are shown in yellow, red, and green, respectively. The first row images are the top-five highest F-score results and the second row images are the top-five lowest F-score results.

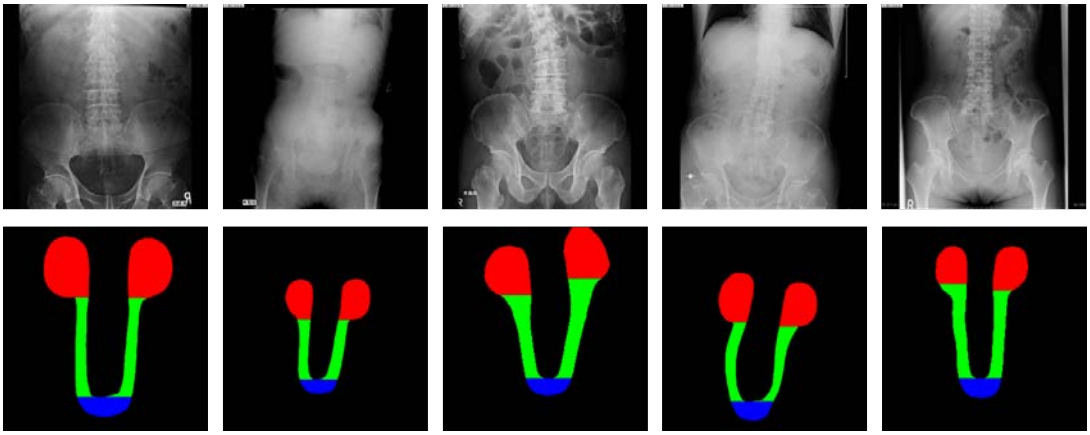


Figure 4.5: Plain abdominal x-ray images (top), and their KUB region maps where kidneys, ureters, and bladder regions are represented in red, green, and blue, respectively.

4.3 Urinary Stones Synthesis Experiments

4.3.1 Experiment setup

We used full abdominal x-ray images consisted of 1,159 I_{sc} and 740 I_{sf} for the urinary stones segmentation experiment. The ground-truth masks of urinary stones were manually drawn by the urology experts for every stone-contained image. We evaluated segmentation performance using five-fold cross-validation. I_{sc} samples were divided into 64% training images, 16% validating images, and 20% testing images in each validation experiment. $G(I_{sc})$ and $G(I_{sf})$ dataset were used only as additional training samples for the network. All dataset for urinary stones segmentation are displayed in Table 4.2. The stones segmentation results were evaluated using pixel-wise and region-wise metrics as described in 4.1.

Table 4.2: Abdominal x-ray dataset for urinary stones segmentation.

	Real	Synthetic		Total
	I_{sc}	$G(I_{sc})$	$G(I_{sf})$	
Train	740	7400	7400	15540
Train per epoch	740	370	370	1480
Validate	185	-	-	185
Test	234	-	-	234

4.3.2 Results and Discussion

Image quality assessment of inpainted images

We evaluated the quality of inpainted images using full-reference image quality assessment (FR-IQA) methods including MSE, PSNR, and SSIM [68], as shown in Table 4.3. Fig. 4.6 illustrates the results of an inpainting network implemented for testing samples. The input images in the stone region were trained to generate both a stone region and its surrounding region in the missing region as illustrated in Fig. 4.6 (columns 1-5), whereas the input images in non-stone regions were trained to fill the missing regions as illustrated in Fig. 4.6 (columns 6-10).

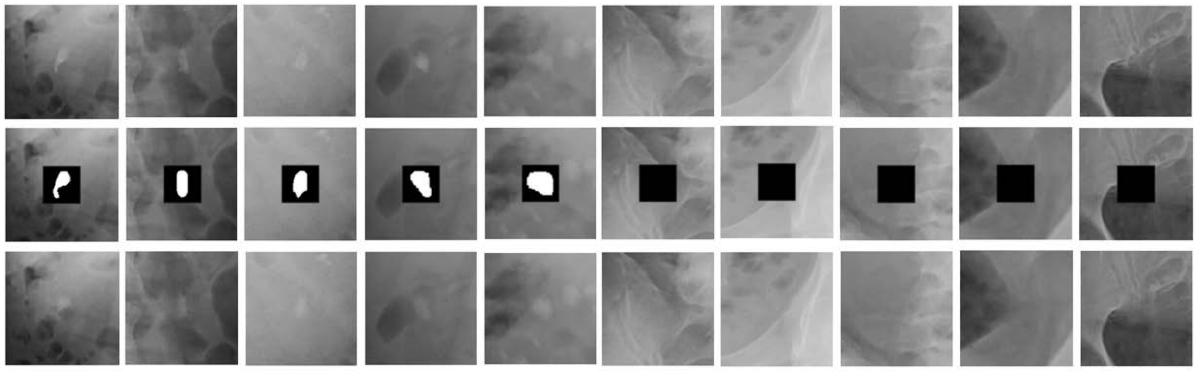


Figure 4.6: Illustration of the original cropped stone region images (1^{st} row), input images for the stone inpainting network(2^{nd} row), and synthesized urinary stone results generated by the stone inpainting network (3^{rd} row).

Table 4.3: Image quality assessment of our inpainted stone and non-stone results.

IQA methods	Testing samples		Average
	Stone mask	Non-stone mask	
MSE	0.00009	0.00007	0.00008
PSNR	42.54418	43.32517	42.93468
SSIM	0.99318	0.99342	0.99330

Segmentation results by the model trained synthetic training data

In this experiment, we compared the pixel-wise and region-wise segmentation results of the MultiresUnet model trained with different training data, namely, real stone-contained (I_{sc}), real stone-free (I_{sf}), synthetic stone-contained ($G(I_{sc})$), and synthetic stone-free ($G(I_{sf})$). The model trained with only I_{sc} was selected as the baseline because its pixel-wise and region-wise F score was superior to that of the model trained with both I_{sc} and I_{sf}. The results in Table 4.4 and 4.5 demonstrate that our proposed synthetic training data could significantly improve segmentation results when compared to a baseline, and the model trained with I_{sc}, $G(I_{sc})$, and $G(I_{sf})$ could achieve the highest scores in all pixel-wise scores and region-wise recall, region-wise F_1, and region-wise F_2 scores. The proposed method outperformed the baseline 2.12% pixel-wise F_1 score (67.47 % to 69.59 %), and 2.13% region-wise F_1 score (66.01 % to 68.14 %). For region-wise evaluation, these synthetic training samples significantly improved recall scores in all experiments; thus, the improvement is obviously seen in region-wise F_2 score, in which FNs are weighted more than FPs. Fig. 4.7 shows the 5-fold cross validation training loss and dice coefficient for a baseline (real) and the proposed method (real+syn.), demonstrating that the proposed method's validation loss was lower than a baseline and its validation dice coefficient was also higher than a baseline.

Additionally, we performed statistical analysis on pixel-wise and region-wise F_1 score results using an independent two-sample t-test comparing the baseline method to those trained with real and synthetic training data, as shown in Fig. 4.8. For pixel-wise evaluation, I_{sc}+$G(I_{sc})$, I_{sc}+$G(I_{sf})$, and I_{sc}+$G(I_{sc})$+$G(I_{sf})$ training data all have a significantly higher F_1 score than the baseline ($p < 0.05$). For region-wise evaluation, MultiResUnet model trained with I_{sc}+$G(I_{sf})$, and I_{sc}+$G(I_{sc})$+$G(I_{sf})$ can improve F_1 score significantly ($p < 0.05$).

From the evaluation results, MultiRes U-Net model has the higher scores in both pixel-wise and region-wise scores than the U-Net model, which has 3-times lower parameters, when using

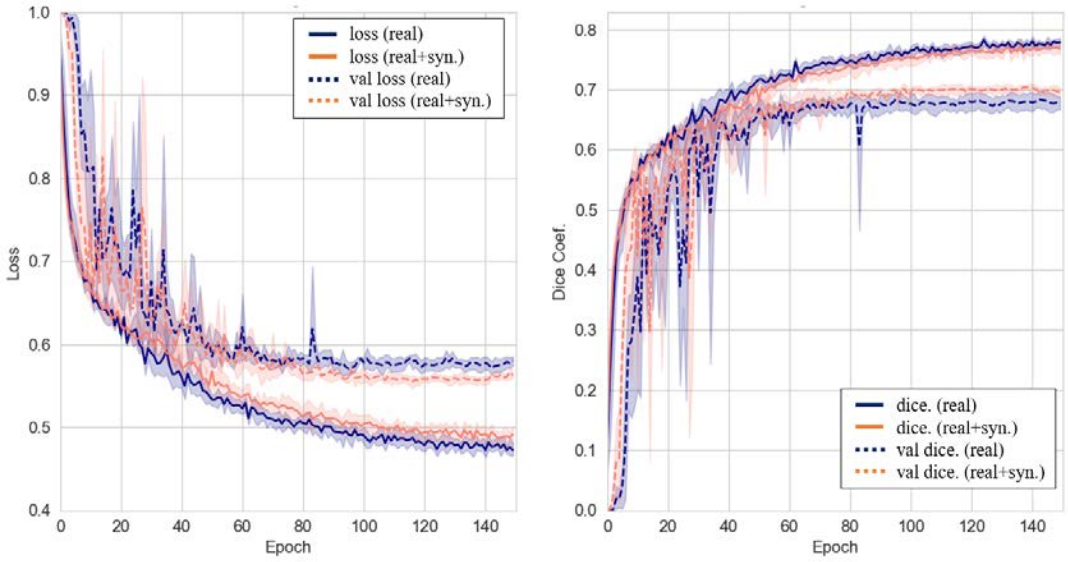


Figure 4.7: The comparisons of training and validation losses (left) and dice coefficients (right) in 5-fold cross validation for the MultiResUnet model trained with different training data.

Table 4.4: Pixel-wise and region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D.%) of the MultiResUnet model trained with different training data.

Training data		Pixel-wise evaluation		
Real	Synthetic	Recall (%)	Precision (%)	F1 score (%)
I_{sc}	-	72.05 (±2.01)	63.05 (±1.76)	67.47 (±0.54)
I_{sc} + I_{sf}	-	71.29 (±0.59)	63.46 (±2.60)	67.13 (±1.67)
I_{sc}	$G(I_{sc})$	72.18 (±1.14)	65.52 (±1.04)	68.68 (±0.60)
I_{sc}	$G(I_{sf})$	72.07 (±0.71)	66.07 (±0.38)	68.97 (±0.34)
I_{sc}	$G(I_{sc})$+$G(I_{sf})$	**72.84 (±1.65)**	**66.65 (±0.97)**	**69.59 (±0.45)**

Table 4.5: Region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D.%) of the MultiResUnet model trained with different training data.

Training data		Region-wise evaluation			
Real	Synthetic	Recall (%)	Precision (%)	F1 score (%)	F2 score (%)
I_{sc}	-	64.11 (±1.92)	68.02 (±3.36)	66.01 (±1.09)	64.86 (±1.19)
I_{sc} + I_{sf}	-	61.99 (±1.31)	66.40 (±3.35)	64.12 (±1.90)	62.82 (±1.38)
I_{sc}	$G(I_{sc})$	65.04 (±0.81)	68.93 (±0.62)	66.93 (±0.36)	65.73 (±0.61)
I_{sc}	$G(I_{sf})$	65.42 (±1.33)	**70.57 (±0.93)**	67.90 (±0.82)	66.39 (±1.10)
I_{sc}	$G(I_{sc})$+$G(I_{sf})$	**66.74 (±1.86)**	69.60 (±1.96)	**68.14 (±1.12)**	**67.29 (±1.45)**

the same training data comparisons. Although, models trained with only synthetic training samples has very low scores, synthetic samples are useful when combining them with the real samples in the training data as shown in the improvements of F_2 score compared to the baseline's results. These synthetic training samples can increase recall score, for all experiments while the precision score is decreased in region-wise evaluation. These results mean that the models trained with real and synthetic samples detect more stone, and also predict more false positive results as the trade-off.

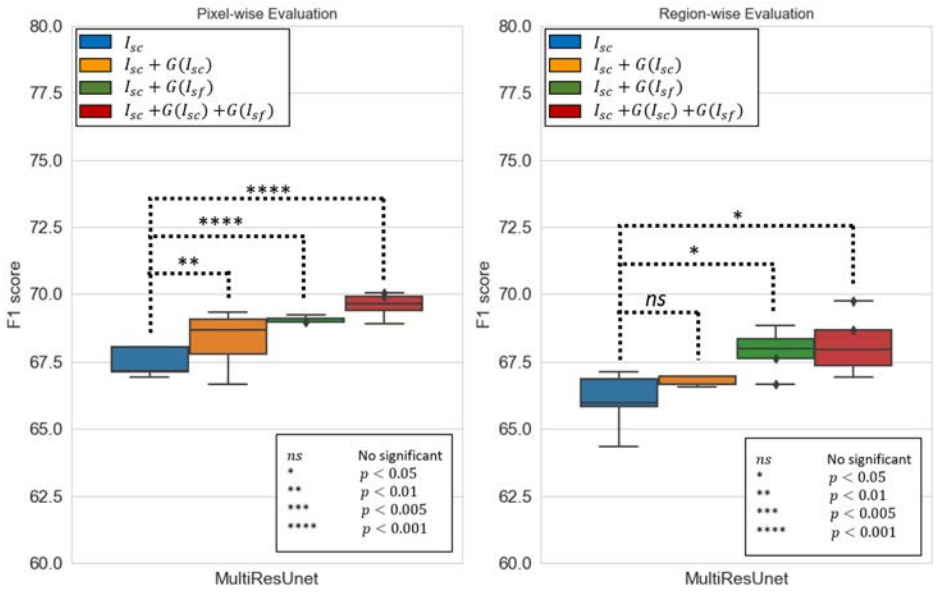


Figure 4.8: The comparison of pixel-wise (left) and region-wise (right) F_1 score of the MultiResUnet model trained with different training data.

Experimental results by Unet-based models

In addition, we compared the state-of-the-art Unet-based models trained with only real stone-contained data (I_{sc}) to those trained with both real stone-contained and all synthetic data (I_{sc}+$G(I_{sc})$+$G(I_{sf})$). The Unet-based models used in this experiment are the original U-Net, ResUnet, Unet++ [**?**], Attention Unet, MultiResUnet , TransUnet, and UTNet. In comparison to other Unet-based models, the MultiResUnet model has the highest recall and F_1 scores for pixel-wise results, and the highest recall, F_1, and F_2 scores for region-wise results. While Unet++ trained on real combined with synthetic samples has the best pixel-wise precision, and the one with only real data has the best region-wise precision. As shown in the pixel-wise and region-wise evaluation results in Table 4.6 and 4.7, all models trained on real data with additional synthetic training data ($G(I_{sc})$ and $G(I_{sf})$) achieved higher pixel-wise F_1 score, region-wise F_1, and F_2 scores than the baselines that was trained with only real data.

Furthermore, we investigated the effect of the stone size on the region-wise recall. All urinary stones were classified according to their size, including small-sized stones (0-200 pixels), medium-sized stones (201-500 pixels), and large-sized stones ($>$ 500 pixels) based on the image's resolution of 1,024 $\times$ 1,024 pixels. The comparison of recalls across different stone size groups in Fig. 4.9 demonstrates that while all baseline models detected large stones well (recall $>$ 0.8), their performance deteriorated significantly for medium and small stones. The addition of synthetic training samples ($G(I_{sc})$, and $G(I_{sc})$) significantly improved the region-wise recall for all models, particularly for small stones, but had a slight effect on recall scores for medium and large stones.

Table 4.6: Pixel-wise and region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D. %) by state-of-the-art Unet-based models trained with different training data.

Model	Training data		Pixel-wise evaluation		
	Real	Syn.	Recall (%)	Precision (%)	F1 score (%)
U-Net	✓	-	71.13 (±1.95)	64.31 (±1.57)	67.51 (±0.43)
U-Net	✓	✓	71.28 (±1.08)	66.6 (±0.88)	68.86 (±0.73)
ResUnet	✓	-	68.13 (±2.37)	66.37 (±1.16)	67.21 (±1.11)
ResUnet	✓	✓	68.40 (±1.02)	68.02 (±1.18)	68.20 (±0.73)
Unet++	✓	-	66.86 (±1.20)	67.19 (±1.67)	67.02 (±1.05)
Unet++	✓	✓	68.02 (±1.48)	**68.74 (±1.58)**	68.35 (±0.09)
Attention Unet	✓	-	70.57 (±0.96)	63.20 (±1.12)	66.67 (±0.48)
Attention Unet	✓	✓	71.29 (±1.27)	64.24 (±0.80)	67.58 (±0.73)
MultiResUnet	✓	-	72.05 (±2.01)	66.07 (±1.76)	67.47 (±0.54)
MultiResUnet	✓	✓	**72.84 (±1.65)**	66.65 (±0.97)	**69.59 (±0.45)**
TransUnet	✓	-	67.83 (±1.25)	60.94 (±2.83)	64.16 (±1.38)
TransUnet	✓	✓	64.79 (±0.54)	69.02 (±1.32)	66.83 (±0.48)
UTNet	✓	-	65.21 (±3.23)	60.10 (±1.93)	62.49 (±1.29)
UTNet	✓	✓	66.46 (±1.97)	61.71 (±1.03)	64.00 (±1.44)

Table 4.7: Region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D. %) by state-of-the-art Unet-based models trained with different training data.

Model	Training data		Region-wise evaluation			
	Real	Syn.	Recall (%)	Precision (%)	F1 score (%)	F2 score (%)
U-Net	✓	-	62.68 (±1.00)	68.83 (±0.82)	65.62 (±0.50)	63.83 (±0.77)
U-Net	✓	✓	64.77 (±1.80)	69.61 (±1.64)	67.10 (±1.21)	65.68 (±1.50)
ResUnet	✓	-	60.05 (±2.69)	71.08 (±1.35)	65.10 (±1.27)	61.98 (±2.16)
ResUnet	✓	✓	61.21 (±1.00)	70.97 (±2.20)	65.73 (±0.93)	62.94 (±0.93)
Unet++	✓	-	58.79 (±1.48)	**71.63 (±2.94)**	64.58 (±1.87)	60.98 (±1.57)
Unet++	✓	✓	61.64 (±1.08)	70.05 (±2.36)	65.58 (±0.81)	63.16 (±0.73)
Attention Unet	✓	-	62.85 (±0.46)	67.91 (±1.33)	65.28 (±0.47)	63.80 (±0.27)
Attention Unet	✓	✓	65.32 (±1.09)	67.65 (±1.13)	66.46 (±0.20)	65.77 (±0.69)
MultiResUnet	✓	-	64.11 (±1.92)	68.02 (±3.36)	66.01 (±1.09)	64.86 (±1.19)
MultiResUnet	✓	✓	**66.74 (±1.86)**	69.60 (±1.96)	**68.14 (±1.12)**	**67.29 (±1.45)**
TransUnet	✓	-	58.14 (±1.96)	61.05 (±5.20)	59.56 (±2.37)	58.70 (±1.60)
TransUnet	✓	✓	57.21 (±0.68)	68.96 (±1.29)	62.53 (±0.68)	59.22 (±0.62)
UTNet	✓	-	58.48 (±3.24)	64.26 (±1.69)	61.24 (±2.33)	59.55 (±2.88)
UTNet	✓	✓	62.49 (±1.27)	64.70 (±3.40)	63.58 (±2.24)	62.92 (±1.63)

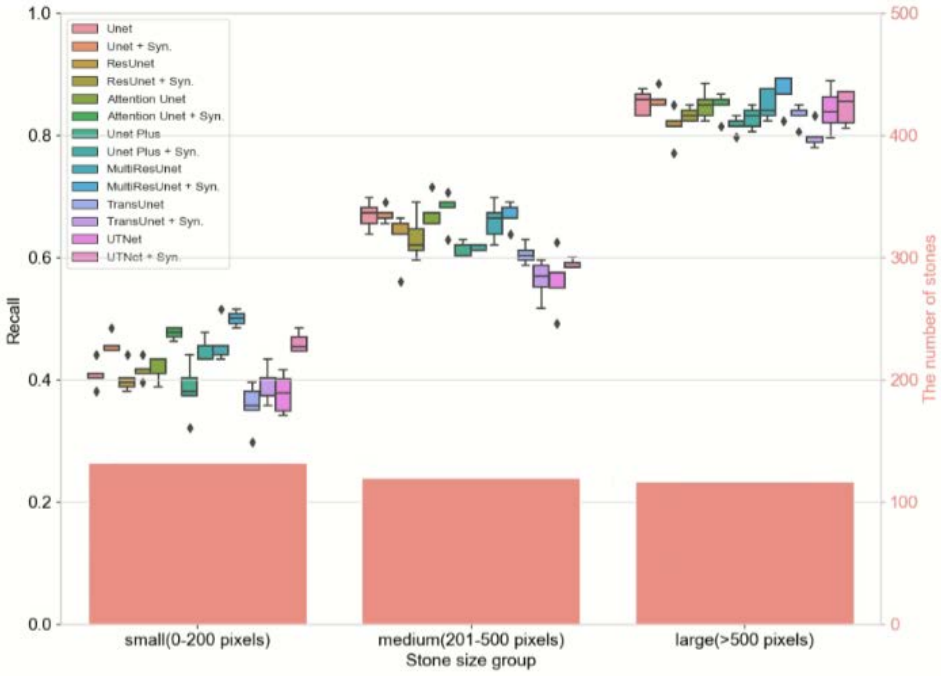


Figure 4.9: The comparison of region-wise recalls of state-of-the-art methods trained with and without synthetic training samples in different stone size groups.

This method, by increasing the number of positive training data $G(I_{sc})$ and $G(I_{sf})$, can support the network in learning to segment urinary stones using a wider variety of images. This method augments the number and variety of positive training samples, which is important when training deep learning to detect urinary stones with irregular shapes, locations, or background properties. Although lower region-wise precision in some models means the model is more likely to predict more FPs when trained with synthetic images, the model also detects more TPs as a trade-off, as evidenced by a significant increase in region-wise F_2 score.

This augmentation method is important for medical imaging applications, where the number of positive cases is typically less than the negative cases. This method is important for medical imaging applications in which the number of positive cases is typically lower than the number of negative cases. By utilizing existing medical images of healthy samples, this method can also be used to reduce the number of actual positive samples required and also improve the segmentation performance of deep learning models.

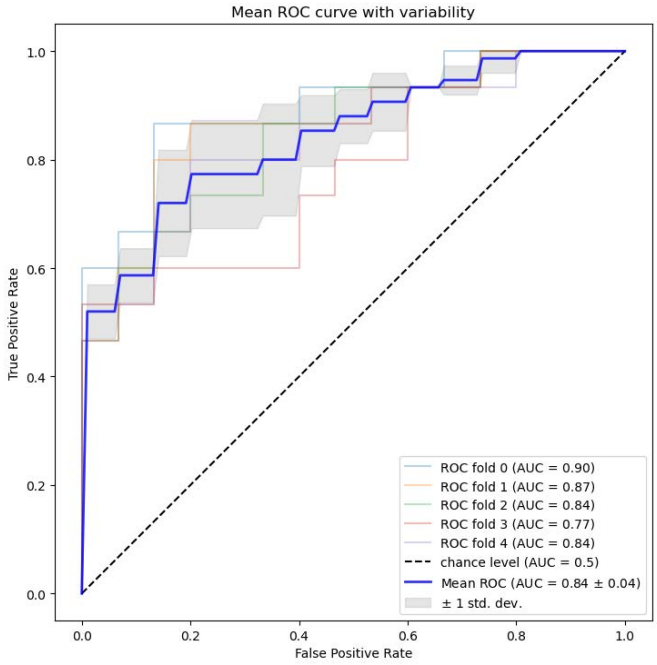


Figure 4.10: Mean ROC curve of bladder stone classification model for 5-fold cross validation.

Table 4.8: Bladder stone classification results measured by recall, precision, and accuracy (average ± S.D.).

Recall Average (±S.D.)	Precision Average (±S.D.)	Accuracy Average (±S.D.)
0.76 (±0.09)	0.83 (±0.03)	0.80 (±0.05)

4.4 False Bladder Stones Detection

The pre-trained VGG16 fine-tuned with our dataset was evaluated using 5-fold cross validation, with 120 cropped-stone images for training and validating, and 30 images for testing, including 15 bladder stone images, and 15 non-bladder stone images. The classification model achieved 0.84 (± 0.04) AUC, as shown in the ROC curve (receiver operating characteristic curve) in Fig. 4.10. The model achieved 0.76 (± 0.09) recall, 0.83 (± 0.03) precision, and 0.80 (± 0.05) accuracy, as shown in Table 4.8.

The detection of false bladder stones was implemented in the post-processing of urinary stones segmentation by 2^{nd} stage network to reduce false positive results. We compared the region-wise segmentation results between the proposed method implementing the proposed false bladder stones detection and the one without it as shown in Table 4.9. Although, this method reduced the region-wise recall (71.84% to 70.36%), it improved the precision (65.55% to 67.76%) and F_1 score (68.55% to 69.03%). The segmentation results comparing between the proposed method with and without post-processing is shown in 4.11, which the proposed post-processing stage can reduce the false prediction in bladder region.

Table 4.9: Comparative stones segmentation results between the proposed method with and without false bladder stone detection measured by region-wise recall, precision, and F_1 score (average ± S.D.%).

False stones detection	Recall (%) Average (±S.D.)	Precision (%) Average (±S.D.)	F_1 score (%) Average (±S.D.)
x	71.84 (±1.42)	65.55 (±2.41)	68.55 (±1.44)
✓	70.36 (±1.18)	67.76 (±1.87)	69.03 (±1.21)

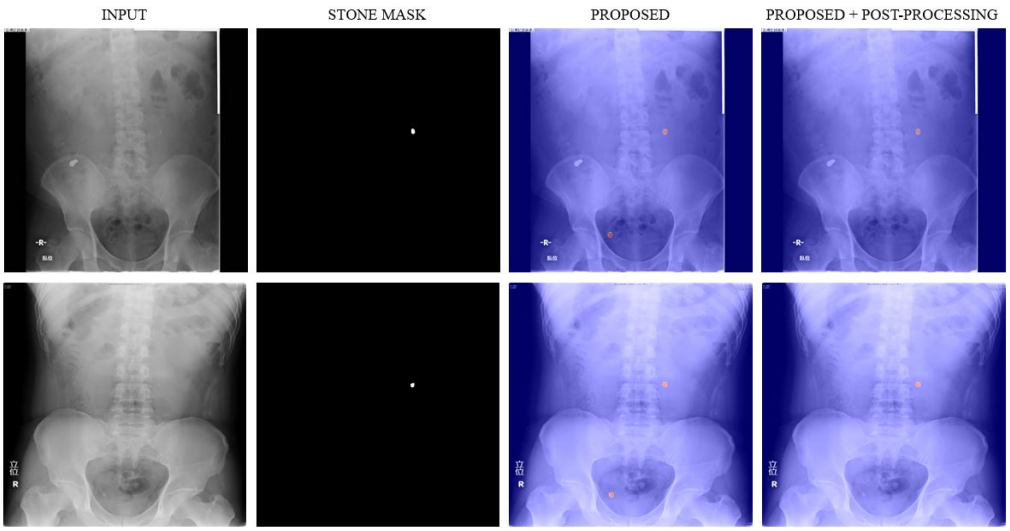


Figure 4.11: The comparison results between proposed method without post-processing and the proposed method implemented post-processing.

Table 4.10: Pixel-wise evaluation of segmentation results (average±S.D.%) by different training methods. The highlight cells represent the scores that difference compared with the baseline are statistically significant ($p < 0.05$).

Method	**Recall (%)**	**Precision (%)**	F_1 **score (%)**	F_2 **score (%)**
Baseline	69.94 (±1.17)	62.94 (±2.26)	66.22 (±0.96)	68.40 (±0.65)
Baseline + iw.	72.29 (±1.72)	61.07 (±1.69)	66.18 (±0.54)	69.70 (±0.92)
Baseline + aug.	71.81 (±0.67)	62.59 (±1.47)	66.87 (±0.85)	69.74 (±1.16)
Baseline + aug. + iw.	74.10 (±1.94)	60.77 (±1.24)	66.76 (±0.93)	70.97 (±1.38)
Baseline + part.	70.42 (±0.86)	**64.57 (±1.77)**	67.36 (±1.29)	69.16 (±1.01)
Baseline + part. + iw.	73.45 (±1.15)	61.87 (±0.56)	67.16 (±0.45)	70.80 (±0.80)
Baseline + part. + aug.	71.72 (±0.33)	64.00 (±0.63)	67.64(±0.30)	70.03 (±0.21)
Proposed	**73.86 (±1.08)**	62.57 (±0.73)	**67.74 (±0.17)**	**71.28 (±0.64)**

4.5 Urinary Stones Segmentation Experiments

4.5.1 Experiment setup

We evaluated urinary stones segmentation performance by implementing the proposed framework using 5-fold cross-validation. Stone-contained (I_{sc}) samples were divided into 64% training images, 16% validating images, and 20% testing images in each validation experiment. Stone-free (I_{sf}) samples were used only in experiments using stone-embedding augmentation. The predicted results were produced by passing the input images to the U-Net model trained with different approaches and then converting the output in the last layer to a binary image using a 0.5 threshold value.

4.5.2 Overall Urinary Stones Segmentation Results

Pixel-wise and region-wise evaluation of urinary stones segmentation results (mean ± s.d.) are presented in Tables 4.10 and 4.11, respectively. For comparison, we chose the U-Net model trained with full real stone-contained images I_{sc} as the baseline method. Based on our pixel-wise and region-wise results, all experiments can outperform the baseline in pixel-wise and region-wise recall and F_2 scores. Our proposed method, the model trained with partitioned real and synthesized images, and implementing inverse weighting map (Partitioned $I_{sc} + G(I_{sc}) + G(I_{sf})$ + iw), achieves the highest recall and F_2 score in both pixel-wise and region-wise evaluation. Although this method produces a lower precision score than the baseline, it significantly improves the recall score as a trade-off, which can outperform the baseline 8.18 % pixel-wise (56.65 % to 64.83 %) and 8.67% region-wise F_2 score (62.19 % to 70.86 %), respectively.

4.5.3 Effect of each proposed method

In this section, we further investigated the effect of each parameter, including input type (full vs. partition), training data (real vs. real + synthetic), and inverse weighting (with vs. without). We evaluated both pixel-wise and region-wise F_2 score and present the statistical analysis based on the independent two-sample t-test between pairs.

Table 4.11: Region-wise evaluation of segmentation results (average±S.D.%) by different training methods. The highlight cells represent the scores that difference compared with the baseline are statistically significant ($p < 0.05$).

Method	**Recall (%)**	**Precision (%)**	F_1 **score (%)**	F_2 **score (%)**
Baseline	61.04 (±1.49)	67.27 (±2.15)	64.00 (±1.05)	62.19 (±1.17)
Baseline + iw.	66.79 (±1.59)	61.85 (±1.34)	64.23 (±0.77)	65.74 (±1.12)
Baseline + aug.	65.26 (±1.15)	66.17 (±2.04)	65.71 (±1.40)	65.44 (±1.18)
Baseline + aug. + iw.	69.10 (±1.88)	61.09 (±1.45)	64.85 (±0.96)	67.33 (±1.36)
Baseline + part.	64.77 (±0.79)	**72.16 (±3.52)**	68.26 (±1.48)	66.12 (±0.69)
Baseline + part. + iw.	68.82 (±0.98)	67.24 (±0.82)	68.02 (±0.60)	68.50 (±0.77)
Baseline + part. + aug.	66.96 (±0.92)	70.76 (±1.17)	68.81 (±0.83)	67.69 (±0.84)
Proposed	**70.36(±1.18)**	67.76 (±1.87)	**69.03 (±1.21)**	**69.82 (±1.08)**

Input Type: Full vs. Partitioned

All experiments of the U-Net model trained with partitioned images demonstrated a significant improvement in pixel-wise and region-wise scores when compared with their paired experiments trained with full image inputs. Instead of receiving entire images as inputs, the second stage U-Net in our cascaded U-Net pipeline processed each partition cropped by KUB region maps. This approach can preserve more information, especially in pixels of small stone, which can be lost during the image scaling and downsampling. Furthermore, the usage of KUB region maps derived from the 1^{st} stage U-Net model can alleviate the imbalance problem between stones and background by removing irrelevant pixels outside urinary tract region.

Training Data: Real vs. Real + Synthetic

Our proposed stone synthetic augmentation reduced the need for actual positive samples and utilized normal images to improve the performance of the deep learning model. When compared to those without this augmentation, the experimental results show that this method significantly improved recall and F_2 score. This method increases the number and variety of positive training samples, which is important for training deep learning models to detect urinary stones in unusual shapes, locations, or background properties. Lower precision results, on the other hand, indicated that the model trained with stone-embedded images was increasingly predicting false positives. This increased false prediction is thought to be due to the fact that some training augmented stones may not appear realistic enough.

Inverse Weighting Map: With vs. Without

The inverse weighting method compensates for the effect of stone size imbalance on loss calculation by multiplying the high weight assigned to small connected components and the low weight assigned to large connected components. Although the precision was decreased when applying this method, the recall was increased as well. These results indicated that the model could detect more stones while also predicting false ones. The results show that using this method with the FTL significantly improved the F_2 score when compared to those without it.

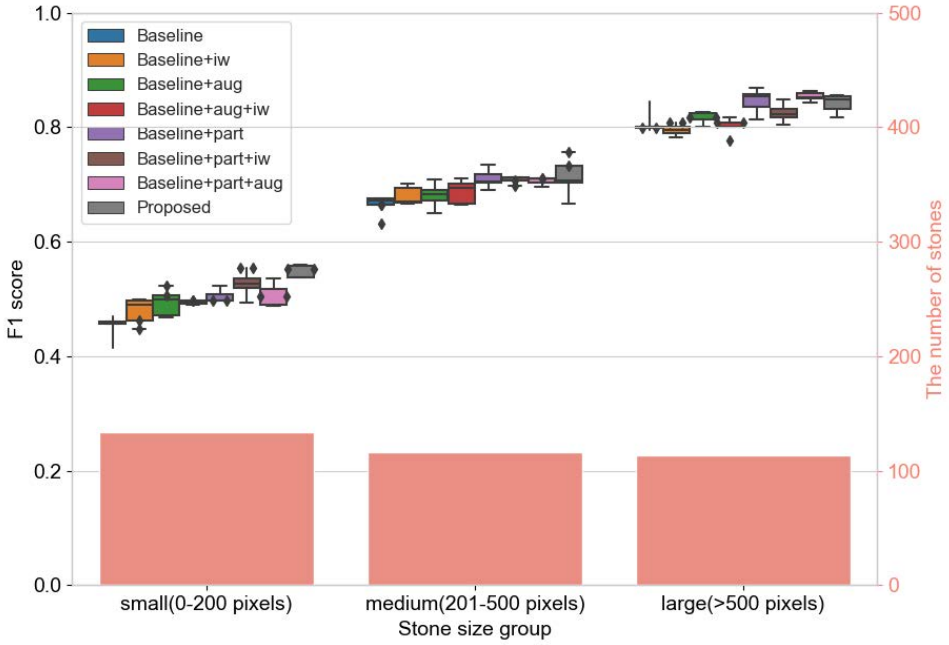


Figure 4.12: The comparison of region-wise F_1 score in different stone size groups.

4.5.4 Stone size vs. region-wise F_1

We also investigated the effect of the stone's size on the segmentation performance (region-wise F_1 score). Firstly, all urinary stones in testing data were classified into 3 categories based on their size, including small-size stones (0-200 pixels), medium-size stones (201-500 pixels), and large-size stones ($>$ 500 pixels), which the image's size is 1,024 $\times$ 1,024 pixels. The result in Fig. 4.12 shows that the region-wise F_1 score was relative to the stone's size, which the larger stones are more detected than the small ones in all experiments. The baseline U-Net model can detect the large stones very well, but the its performance dramatically decreases in small and medium stones. This result also indicated that U-Net model implemented all proposed method (Proposed.) could significantly enhance F_1 score, particularly for small-sized and medium-sized stones, which produced the highest F_1 score in these categories.

4.5.5 Anatomical region of the stone vs. region-wise F_1

For evaluating the stone in different anatomical regions and region-wise F_1 score, we separated all urinary stones into 3 categories based on their location, including kidneys, ureters, and bladder, by using the KUB region maps. The result in Fig. 4.13 shows that stone detection performance decreases significantly in the bladder region, which has the lowest number of stones. The results of stones in ureters region shows that the experiments using partitioned inputs by a two-stage framework significantly produce higher recall scores than the ones using full image inputs. The result of stones in the bladder region shows that our proposed method produces the highest recall score compared to others. Additionally, we investigated the recall scores in 9 categories based on their stone-size and stone-region groups as shown in Fig.4.14. The comparison in this figure shows that the proposed method can improve the stone detection performance in almost every categories. Particularly in difficult categories, our proposed method can achieve higher than 33.5% (9.5% to 43%) in small-bladder category, and 21% (41% to 62%) small-ureters category, produced by the baseline method.

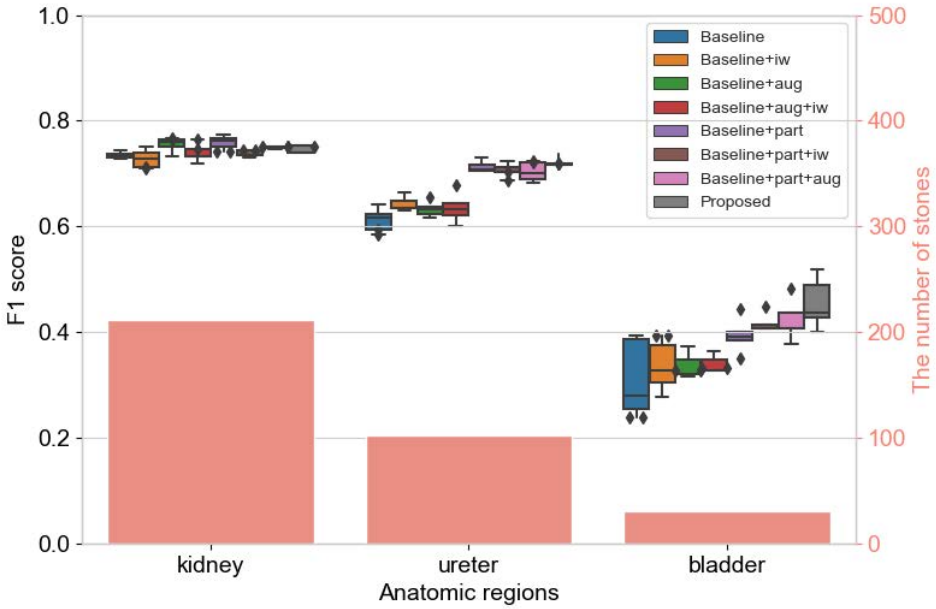
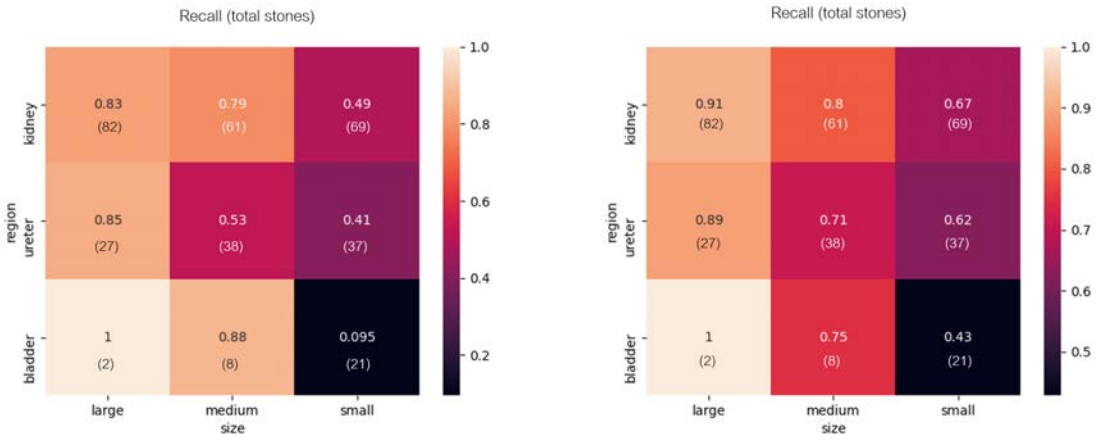


Figure 4.13: The comparison of region-wise F_1 score in different anatomic regions.



(a) (b)

Figure 4.14: Recall results of baseline U-Net model (a), and U-Net model with proposed method (b).

4.5.6 Qualitative Comparison

The illustration shown in Fig.4.15 is a comparison between urinary segmentation results, which the baseline method and our proposed method can detect all stones in images correctly. The two-stage pipeline of our proposed method, which uses partitioned input, can produce a little more precise segmentation results than the baseline, which uses the full image input. Therefore, from these example results, the pixel-wise score from the proposed method are a little better than the baseline, while the region-wise scores from both methods are equal.

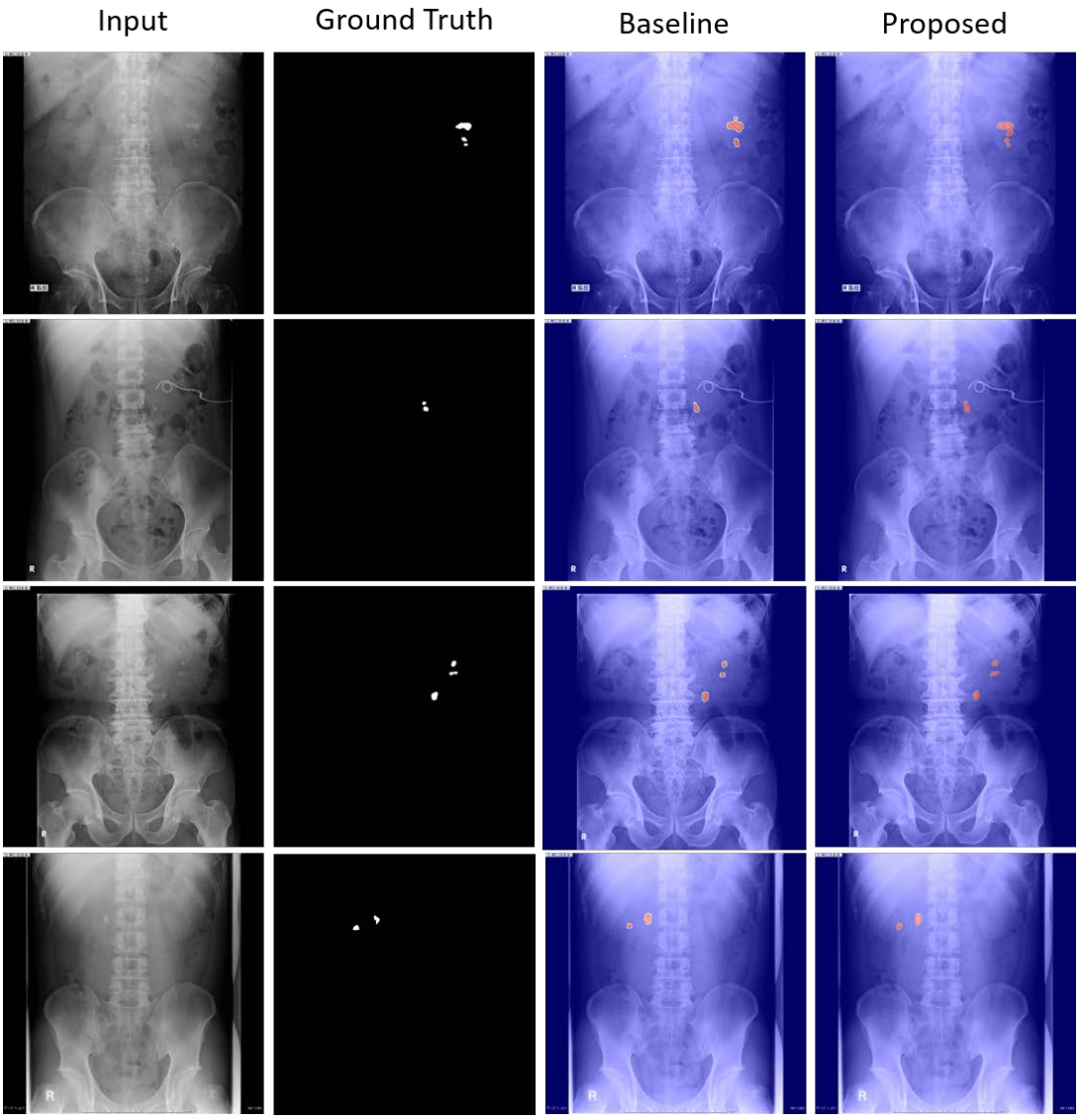


Figure 4.15: Illustration of a comparison between urinary stone segmentation results by a baseline method and those by our proposed method (the heatmap visualization displays predicted stone regions), which both methods can detect all stones.

The illustration shown in Fig.4.16 is a comparison between urinary segmentation results, which our proposed method can detect all stones in images correctly, while some stones are missed in the results by a baseline method. Fig.4.14 shows that our proposed method can detect the stone in difficult region better than the baseline method. For examples, the stone near bone structure can be detected only in our segmentation result. The stones shaded by a bowel gas loop (Fig.4.14 (row 2-4)) are missed in the baseline method, while those by our method can be

detected.

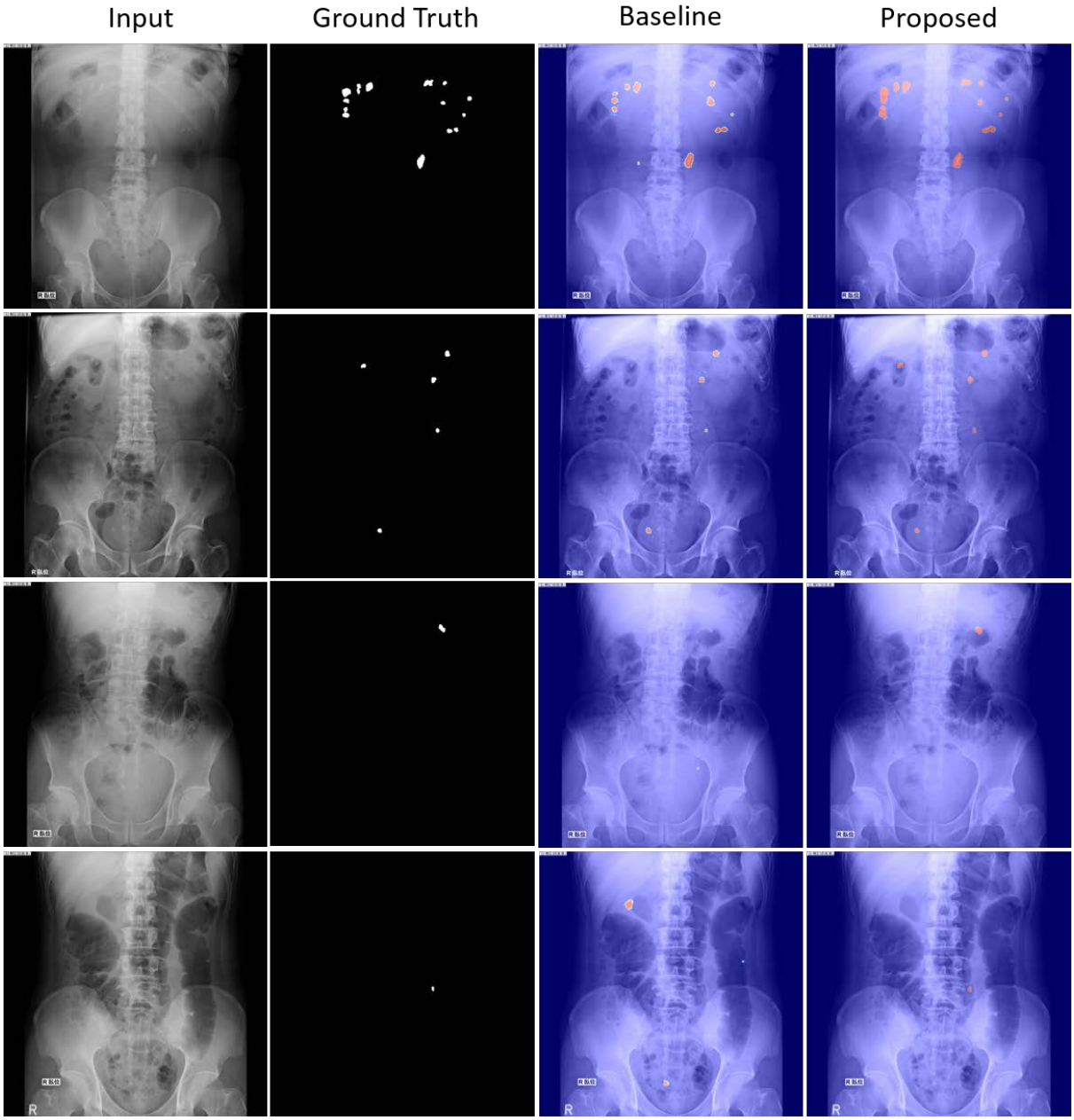


Figure 4.16: Illustration of a comparison between urinary stone segmentation results by a baseline method and those by our proposed method (the heatmap visualization displays predicted stone regions), which the proposed methods can detect all stones, while some stones are missed by a baseline method.

The illustration shown in Fig.4.17 is a comparison between urinary segmentation results, which our proposed method can detect all small stones (area < 200 pixels) in images correctly, while some small stones are missed in the results by a baseline method. Our proposed method, using partitioned input and implementing inverse weighting map for loss calculation, can detect the small stones better than the baseline method.

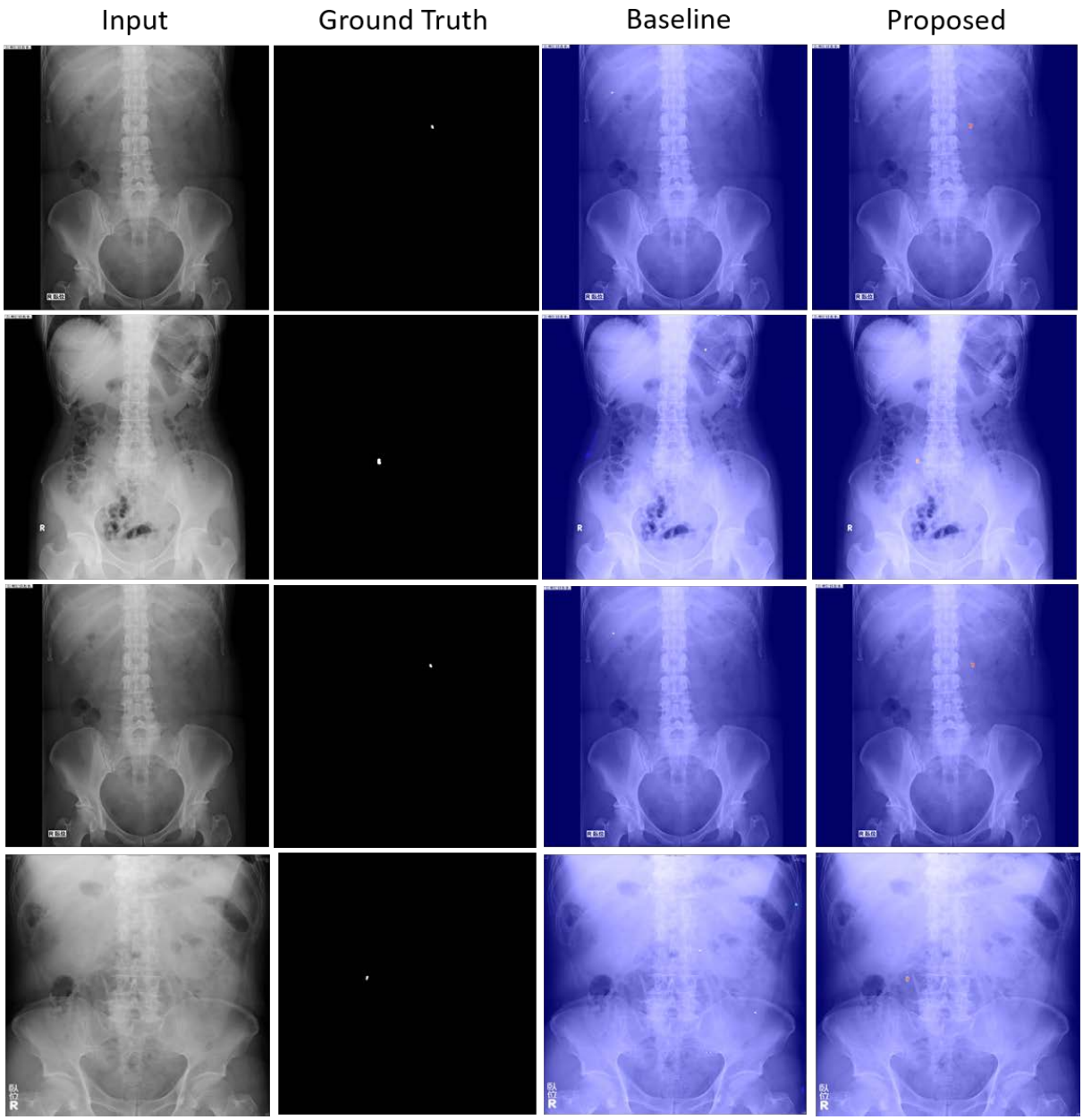


Figure 4.17: Illustration of a comparison between urinary stone segmentation results by a baseline method and those by our proposed method (the heatmap visualization displays predicted stone regions), which the proposed methods can detect all stones, while all small stones are missed by a baseline method.

The illustration shown in Fig.4.18 is a comparison between urinary segmentation results, which our proposed method can detect all stones in ureters and bladder region correctly, while some stones in these region are missed in the results by a baseline method. The stone detection in these region is difficult and usually missed because of the small training data in these case. Furthermore, the stones in ureters usually small and overlapped with the bone structure (pelvic bone) or very closed to spine. Our proposed method, using partitioned input, implementing inverse weighting map, and training with synthetic samples, can improve the segmentation results in this case.

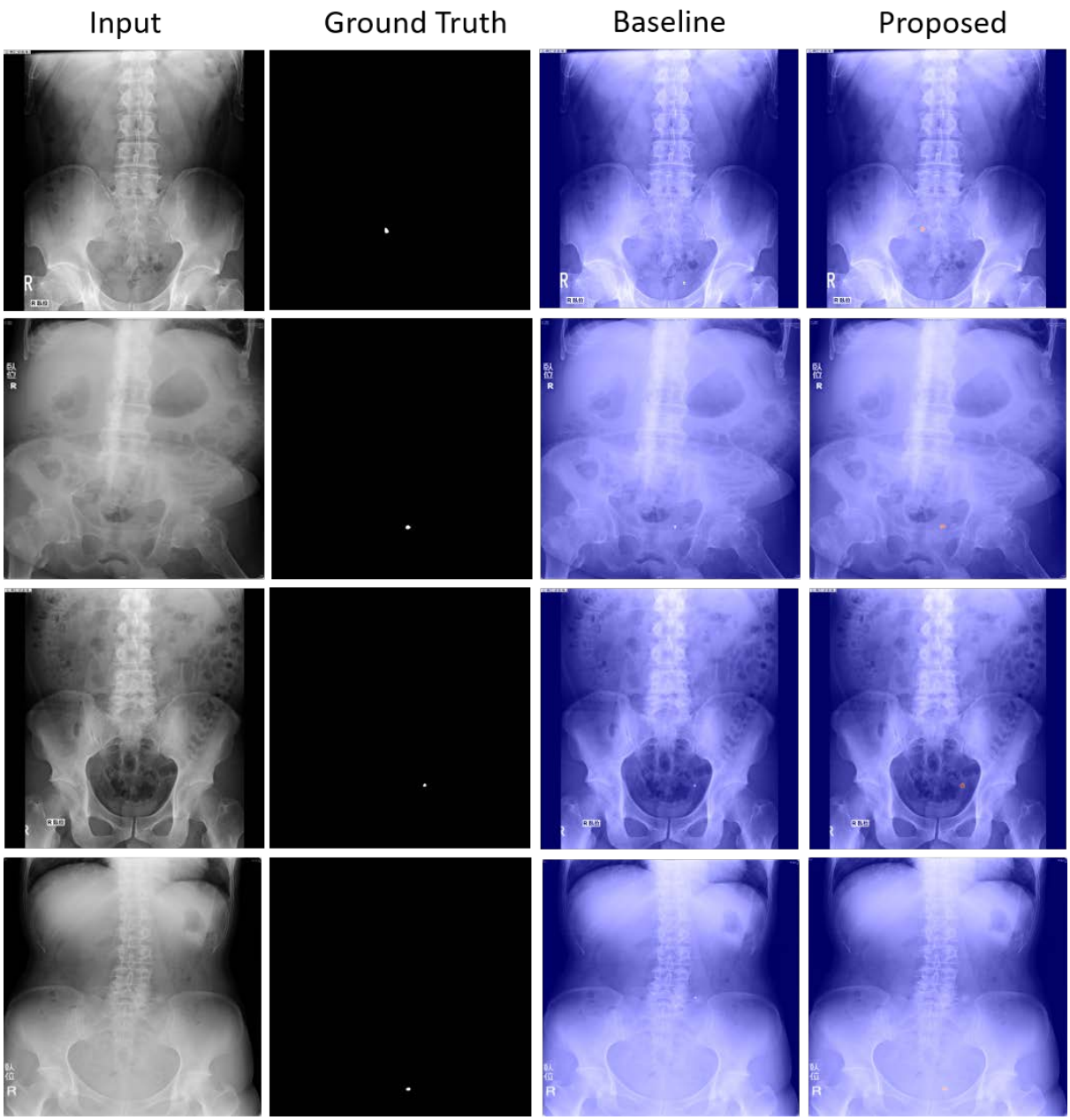


Figure 4.18: Illustration of a comparison between urinary stone segmentation results by a baseline method and those by our proposed method (the heatmap visualization displays predicted stone regions), which the proposed methods can detect all stones in ureters and bladder region, while all stones in these regions are missed by a baseline method.

Although the overall metrics indicated that our proposed method has a better performance for stones detection than a baseline, there are some case, which a baseline can perform better than ours. The illustration shown in Fig.4.19 is a comparison between urinary segmentation results, showing the testing images which the baseline method can detect stones more correctly than our proposed method.

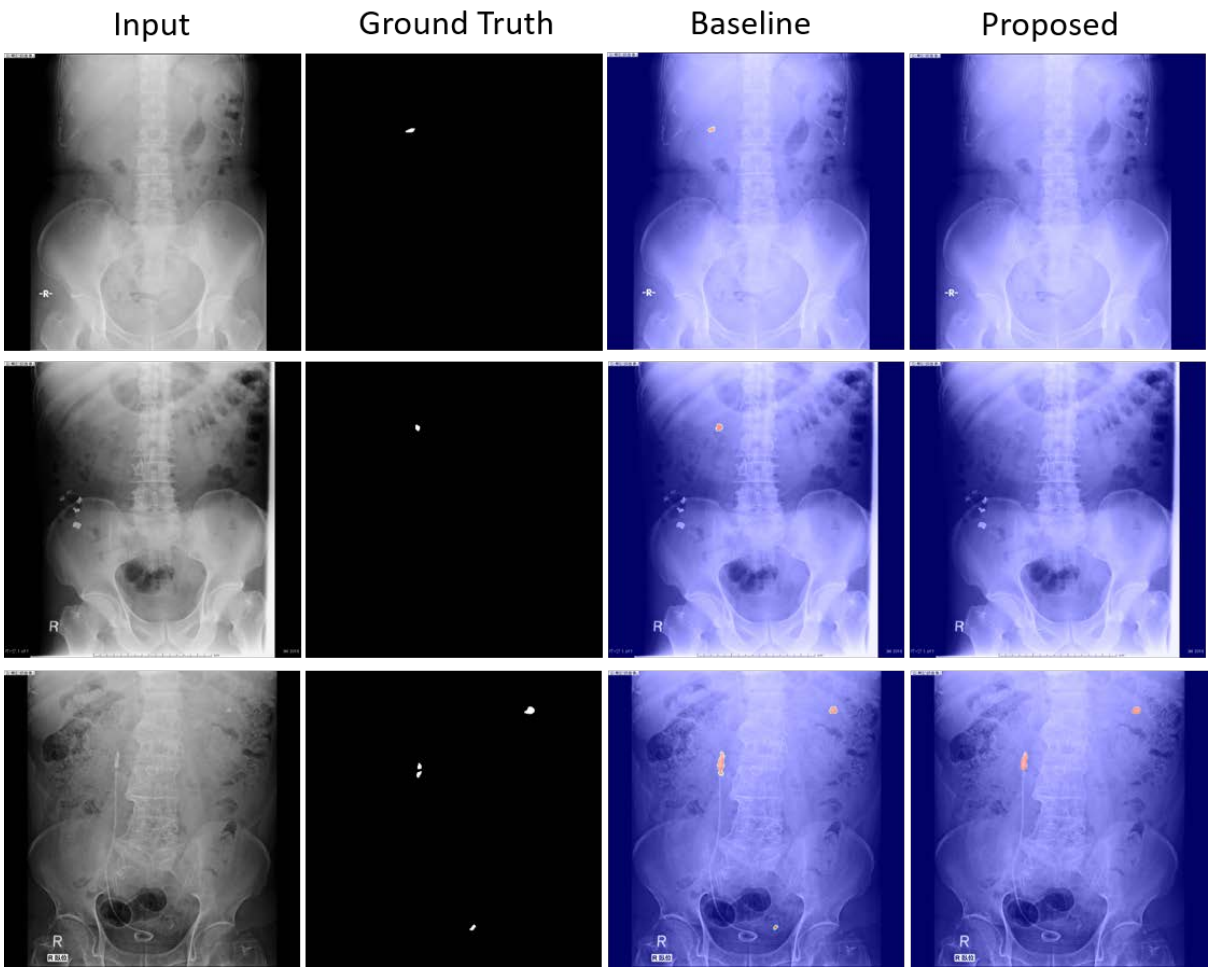


Figure 4.19: Illustration of a comparison between urinary stone segmentation results by a baseline method and those by our proposed method (the heatmap visualization displays predicted stone regions), which a baseline method can perform better than ours.

The illustration shown in Fig.4.20 is a comparison between urinary segmentation results, which the baseline and our proposed method can not detect the stone in these images. These false negative results usually occur in small stones and stones in bladder region. Although some stone such as the one in Fig. 4.21 (3_{rd} row) is not a small stone, it is hardly visible, which is difficult to be detected by the model.

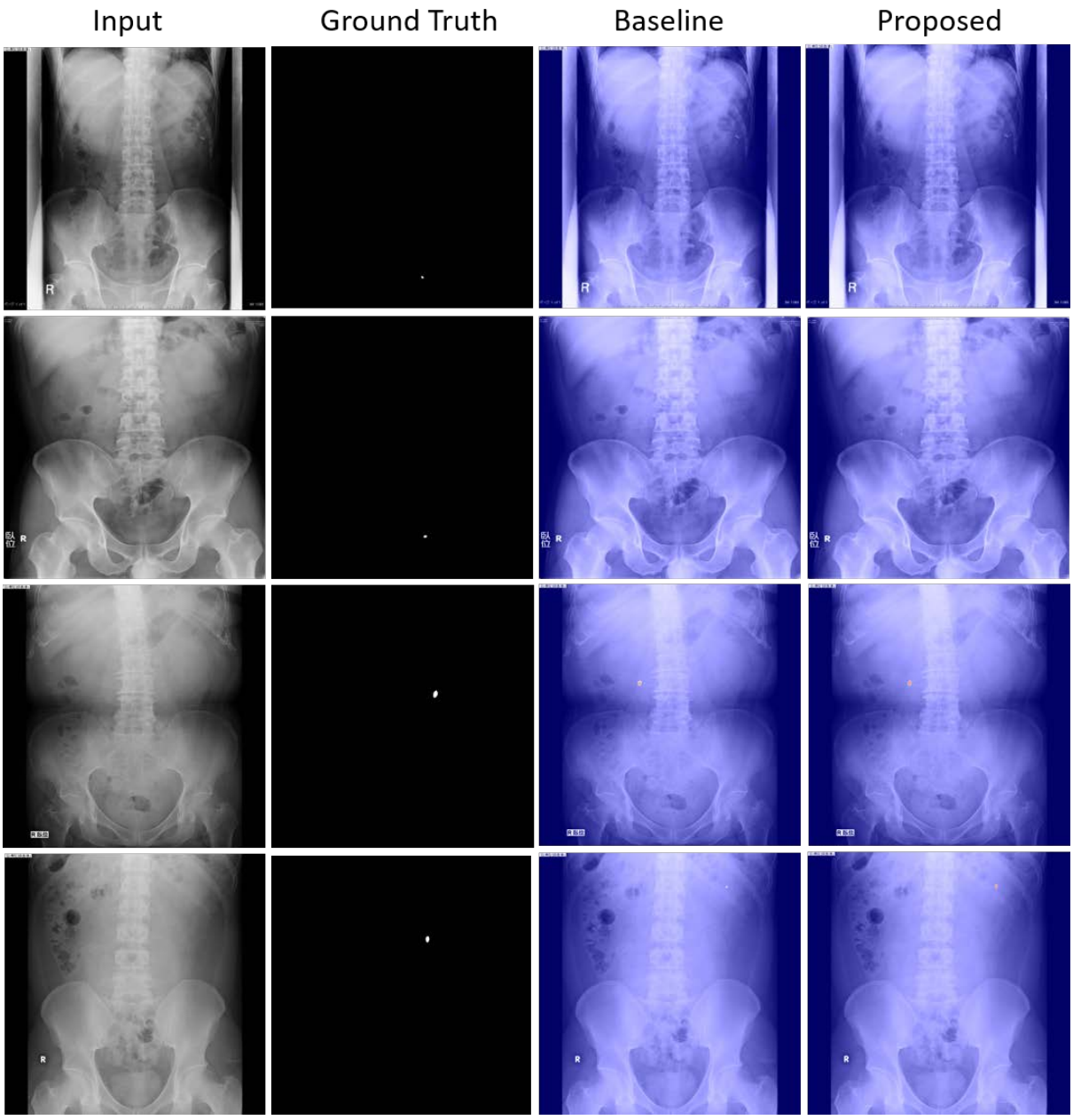


Figure 4.20: Illustration of example false negative results by a baseline method and those by our proposed method.

The illustration shown in Fig.4.21 is a comparison between urinary segmentation results, showing false positive case. The lower region-wise precision score of our proposed method indicates when the model can predict more stone, the false positive results are increased as well.

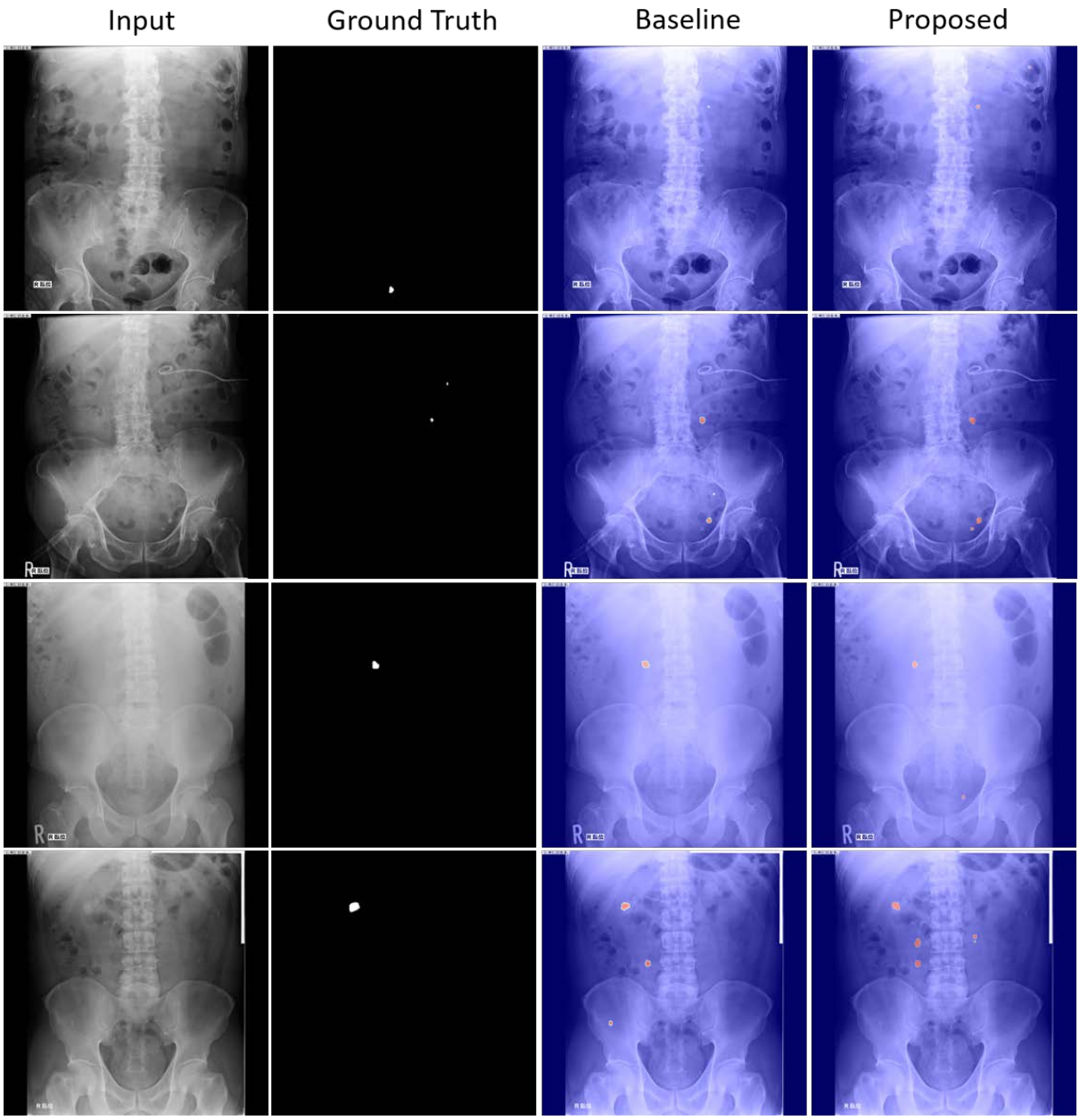


Figure 4.21: Illustration of example false positive results by a baseline method and those by our proposed method.

4.5.7 Comparisons with State-of-the-art Deep Learning Model

In this section, we compared U-Net-based models implemented with our proposed method (partitioned input + stone-embedding augmentation + inverse weighting map) and the baseline U-Net-based models, which were trained using full images without any proposed method. The U-Net variants that we experimented included U-Net, ResUnet, Unet++, Attention Unet, MultiResUnet, and TransUnet models.

The pixel-wise and region-wise evaluation results are shown in Table 4.12 and 4.13, respectively. Unet++ model with the proposed methods has the highest pixel-wise F-score, while MultiResUnet model has the highest region-wise F-score for both baseline approach and the one employing the proposed methods. Plain U-Net model implementing the proposed methods

Table 4.12: Pixel-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D. %) by Unet-based models with and without our proposed pipeline.

Model	Pixel-wise evaluation		
	Recall (%)	Precision (%)	F_1 score (%)
Unet	69.94 (±1.17)	62.94 (±2.26)	66.22 (±0.96)
Unet w/ proposed.	73.86 (±1.08)	62.57 (±0.73)	67.74 (±0.17)
ResUnet	68.42 (±1.03)	63.23 (±0.93)	65.72 (±0.36)
ResUnet w/ proposed.	70.94 (±1.50)	64.27 (±1.25)	67.43 (±0.64)
Attention Unet	67.91 (±2.90)	61.62 (±1.57)	64.59 (±1.80)
Attention Unet w/ proposed.	**73.90 (±0.67)**	60.90 (±1.12)	66.76 (±0.62)
Unet++	68.47 (±1.13)	63.35 (±1.62)	65.79 (±0.84)
Unet++ w/ proposed.	69.00 (±1.85)	**66.87 (±1.91)**	**67.88 (±0.57)**
MultiResUnet	72.71 (±1.24)	62.29 (±3.12)	67.05 (±1.49)
MultiResUnet w/ proposed.	72.64 (±1.62)	63.57 (±1.18)	67.79 (±0.71)
TransUnet	67.83 (±1.25)	60.94 (±2.83)	64.16 (±1.38)
TransUnet w/ proposed.	67.86 (±2.59)	65.02 (±1.28)	66.39 (±1.41)

Table 4.13: Region-wise evaluation of segmentation results measured by recall, precision, and F_B score (average ± S.D. %) by Unet-based models with and without our proposed pipeline.

Model	Region-wise evaluation			
	Recall (%)	Precision (%)	F_1 score (%)	F_2 score (%)
Unet	61.04 (±1.49)	67.27 (±2.15)	64.00 (±1.05)	62.19 (±1.17)
Unet w/ proposed.	70.36 (±1.18)	67.76 (±1.87)	69.03 (±1.21)	69.82 (±1.08)
ResUnet	60.00 (±1.15)	66.65 (±2.08)	63.15 (±1.05)	61.22 (±0.96)
ResUnet w/ proposed.	68.22 (±1.03)	69.51 (±1.70)	68.86 (±0.41)	68.47 (±0.54)
Attention Unet	61.37 (±1.56)	62.89 (±2.77)	62.12(±1.77)	61.67 (±1.51)
Attention Unet w/ proposed.	**72.00 (±1.14)**	62.81 (±2.35)	67.09 (±1.47)	69.95 (±1.09)
Unet++	59.18 (±0.95)	64.90 (±1.80)	61.91 (±1.21)	60.24 (±1.00)
Unet++ w/ proposed.	65.04 (±1.82)	**71.21 (±2.78)**	67.98 (±1.34)	66.19 (±1.43)
MultiResUnet	65.15 (±0.41)	66.72 (±2.06)	65.93 (±0.98)	65.46 (±0.47)
MultiResUnet w/ proposed.	71.34 (±1.35)	67.32 (±1.78)	**69.27 (±0.43)**	**70.50 (±0.72)**
TransUnet	58.14(±1.96)	61.05 (±5.20)	59.56 (±2.37)	58.70 (±1.60)
TransUnet w/ proposed.	65.10 (±1.31)	67.50 (±2.42)	66.28 (±1.27)	65.56 (±1.10)

has better pixel-wise and region-wise F-scores than other baseline U-Net-based variants. Overall, our proposed pipeline can significantly improve F-scores in both pixel-wise and region-wise evaluations as shown by the improvement when compared to the baselines of all Unet-based models.

4.5.8 Limitations and Suggestions for the Future Works

Although the stone-region evaluation shows that our proposed method can detect the stones in the kidney region very well (recall ≈ 0.80), there is some case that the model cannot detect the stones. Based on our stone-size and stone-region evaluations, the small stones in the lower ureters or bladder region are the most challenging case that shows the lowest recall results compared with other cases. An example of this case is demonstrated in Fig.4.22 which our model cannot detect the small stone that is poorly visible in the bladder region. Furthermore, the the experiments implementing stone-synthesized augmentation produced more false-positive

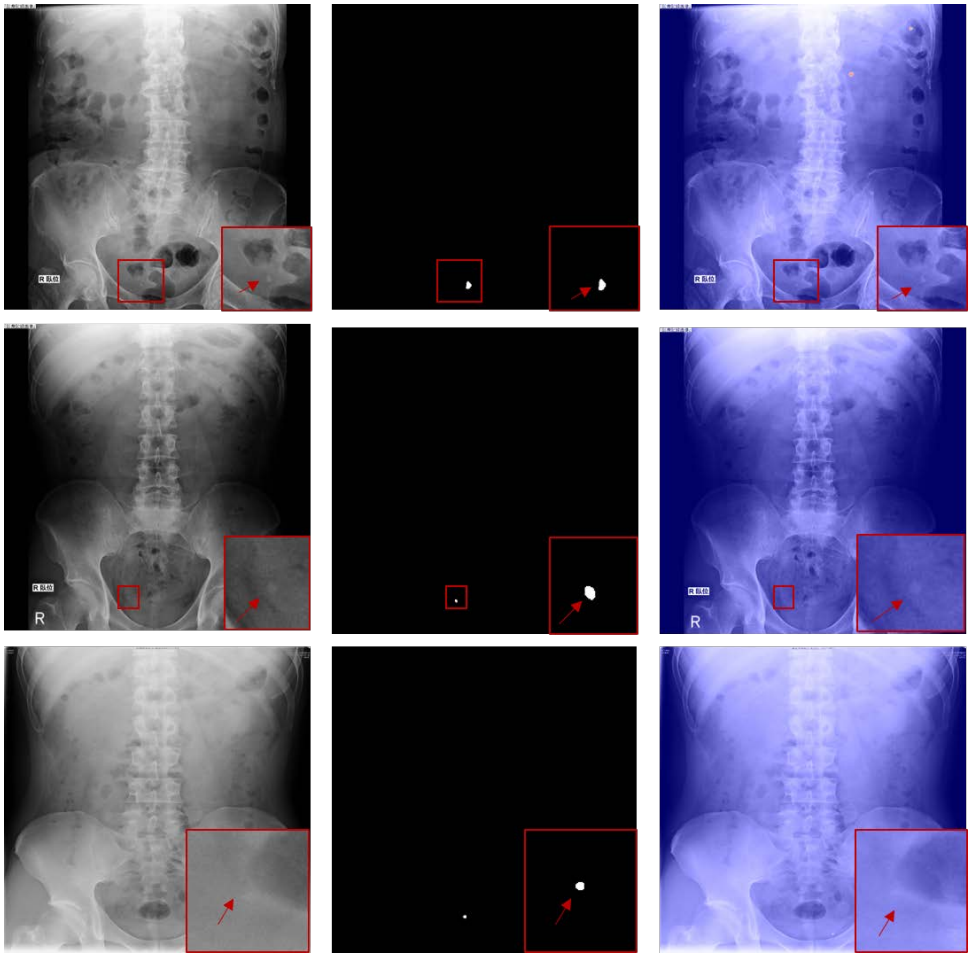


Figure 4.22: False-negative examples by our method (the heatmap visualization displays predicted stone regions). Red boxes show enlarged regions containing urinary stones in bladder region that were missed.

results, hence the precision score became decreased. The reason is that some stone-synthesized images used to train the segmentation network do not look realistic enough as shown in Fig. 4.23. The 1^{st} and 2^{nd} column images display the case that stones are overlapped the medical instrument. As well, 3^{rd} - 5^{th} column images display the case that stones are overlapped other anatomical structure.

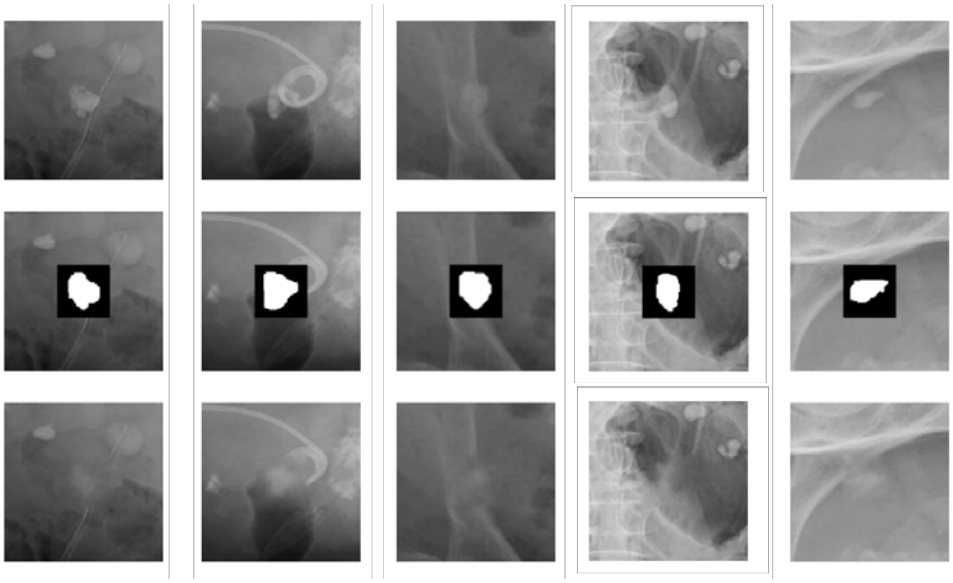


Figure 4.23: The illustration of original cropped stone region images (1^{st} row images), input images for stone inpainting network (2^{nd} row images), and synthetic stone results in failed cases (3^{rd} row images).

Chapter 5

Conclusion

We proposed a two-stage pipeline for automatically segmenting urinary stones in abdominal x-ray images. The first stage model takes the full images as the inputs and generates the KUB region maps representing the approximate locations of urinary organs where stones are present. The second stage model is trained with actual stone-contained images and stone-synthesized images to segment urinary stones from the partitioned images generated cropped by corresponding KUB region maps. The urinary stones segmentation network in the cascaded framework, which processed partitioned images instead of full images, could improve segmentation results by reducing the class imbalance problem and processing images at higher resolution. GAN-based synthetic stone augmentation was presented to increase the number and variety of positive training samples by implementing them with the healthy samples, which usually outnumber those with findings. This augmentation method can improve the performance, particular for stones in rare locations. Our stone-size re-weighting approach used with the focal Tversky loss could significantly improve the performance of detection particularly for small stones. Lastly, the bladder stone classification network also reduced false positive results in the bladder region. The proposed method using the U-Net model produced a 71.28% pixel-wise F_2 score and a 69.82% region-wise F_2 score, which were higher than 2.88% and 7.63% produced by the baseline U-Net model, respectively. The experimentation with other U-Net-based segmentation models also showed that our proposed pipeline could improve F_2 score significantly. In the future, we hope that this automated segmentation of urinary stones in abdominal x-ray images can be used to reduce the burden on screening, and support medical doctors in diagnosis, and treatment planning.

PUBLICATIONS

[1] W. Preedanan *et al.*, "Improvement of Urinary Stone Segmentation Using GAN-Based Urinary Stones Inpainting Augmentation," in *IEEE Access*, vol. 10, pp. 115131-115142, 2022, , doi: 10.1109/ACCESS.2022.3218444

[2] W. Preedanan *et al.*, "Urinary Stones Segmentation in Abdominal X-Ray Images Using Cascaded U-Net Pipeline with Stone-Embedding Augmentation and Lesion-Size Reweighting Approach," in *IEEE Access*, vol. 11, pp. 25702-25712, 2023, doi:10.1109/ACCESS.2023.3257049.

Bibliography

L. Giannossi and V. Summa, “A review of pathological biomineral analysis techniques and classification schemes,” in An Introduction to the Study of Mineralogy, C. Aydinalp, Ed., InTechOpen, InTech, IMAA-CNR, Italy, 2012.

Melissa Conrad Stöppler, 2021, MedicineNet, (https://www.medicinenet.com/kidney_stones/article.htm)

S. R. Khan *et al.*, “Kidney stones,” *Nat. Rev. Dis. Primers*, vol. 2, 2016, p. 16008.

T. Alelign, and B. Petros, “Kidney Stone Disease: An Update on Current Concepts,” *Adv. Urol.*, 2018, pp.1-12.

W. Brisbane, M. R. Bailey, and M. D. Sorensen, “An overview of kidney stone imaging techniques,” *Nat. Rev. Urol.*, vol. 13(11), pp.654-662, doi: 10.1038/nrurol.2016.154.

P. R. Tamilselvi and P. Thangraj, "Segmentation of calculi from Ultrasound kidney images by region indicator with contour segmentation method", global journal of computer science and technology, vol. 11, no. 22, pp. 43-51, Dec. 2011

R. Raja and J. Ranjani, "Segmentation based detection and quantification of kidney stones and its symmetric analysis using texture properties based on logical operator with US images", international journal of computer applications (0975–88867), pp. 8-15.

P. R. Tamilselvi, "Segmentation of renal calculi using Squared Euclidian Distance method", international journal of scientific engineering and technology, vol. 2, no. 7, pp. 651-655.

W. Mahani and Eko Supriyanto, "Comparative evaluvation of ultrasound kidney image enhancement techniques", international journal of computer applications (0975–8887, vol. 21, no. 7, pp. 15-19, May 2011.

W. Mahani and Eko Supriyanto, "Automatic ROI Generation for Kidney Ultrasound Images", Recent Researches in Applied Informatics and Remote Sensing, pp. 71-75.

P. R. Tamilselvi, "Detection of renal calculi using semi automatic segmentation approach", international journal of engineering and innovative technology (IJESIT), vol. 2, no. 3, pp. 547-552, May 2013.

K. Viswanath and R. Gunasundari, “Design and analysis performance of kidney stone detection from ultrasound image by level set segmentation and ANN classification,” *Proc. 2014 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2014*, 2014, pp. 407–414.

P. R. Tamilselvi and P. Thanraj, "computer aided diagnosis system for stone detection and early stone detectionn of kidney stones", *J. Comput. Sci.*, vol. 7(2), 2021, pp. 250-254.

https://doi.org/10.3844/jcssp.2011.250.254.

H. Dave, V. Patel, J. N. Mehta, S.Degadwala, and D. Vyas, "Regional Kidney Stone Detection and Classification In Ultrasound Images," Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021.

P. T. Akkasaligar, S. Biradar and V. Kumbar, "Kidney stone detection in computed tomography images," *2017 Int. Conf. Smart Technol. Smart for Nation (SmartTechCon)*, Bangalore, 2017, pp. 353-356.

L. Y. Myint, S. S. Maung, and K. T. Zar, "Removal of Unwanted Object in 3D CT Kidney Stone Images and 3D Visualization," 24th International Computer Science and Engineering Conference (ICSEC), Bangkok, Thailand, 2020.

O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

O. Ronneberger, P. Fischer, and T. Brox, "The Importance of Skip Connections in Biomedical Image Segmentation," 2016, *arXiv:1608.04117*. [Online]. Available: https://arxiv.org/abs/1608.04117

Z. Zhang, Q. Liuand, and Y. Wang, "Road Extraction by Deep Residual U-Net," *Geoscience and Remote Sensing Letters IEEE*, vol.5, 2018.

Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," 2018, *arXiv:1807.10165*. [Online]. Available: https://arxiv.org/abs/1807.10165

O.Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

N. Ibtehaz1 and M. S. Rahman, "MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," 2019, *arXiv:1902.04049*. [Online]. Available: https://arxiv.org/abs/1902.04049

A. Vaswani *et al.*, "Attention Is All You Need," 2017, *arXiv:1706.03762* . Online]. Available: https://arxiv.org/abs/1706.03762

J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," 2021, *arXiv:2102.04306*. Online]. Available: https://arxiv.org/abs/2102.04306

A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, *arXiv:2010.11929*. Online]. Available: https://arxiv.org/abs/2010.11929

Y. Gao, M. Zhou, and D. Metaxas, "UTNet: a Hybrid Transformer Architecture for Medical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, p.61-71, 2021.

H. Cao *et al.* "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," 2020, *arXiv:2105.05537*, [Online]. Available: https://arxiv.org/abs/2105.05537

M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *J. Magn. Reson. Imaging*, vol. 49, no. 4, 2019, pp. 939–954, doi: 10.1002/jmri.26534.

X. Li *et al.*, "Three-dimensional simulation of lung nodules for paediatric multidetector array CT," *Brit. J. Radiol.*, vol. 82, no. 977, 2009, pp. 401–411.

A. Rashidnasab *et al.*, "Simulation and assessment of realistic breast lesions using fractal growth models," *Phys. Med. Biol.*, vol. 58, no. 16, 2013,pp. 5613–5627.

M. S. Vaz, Q. Besnehard, and C. Marchessoux, "3D lesion insertion in digital breast tomosynthesis images," *Proc. SPIE*, vol. 7961, Mar. 2011, pp. 79615Z-1–79615Z-10.

R. D. Ambrosini and W. G. O'Dell, "Realistic simulated lung nodule dataset for testing CAD detection and sizing," *Proc. SPIE*, vol. 7624, Mar. 2010, p. 76242.

A. P. Peskin and A. A. Dima, "Modeling clinical tumors to create reference data for tumor volume measurement," *Adv. Vis. Comput. (ISVC) 2010*, Lecture Notes in Computer Science, vol 6454. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17274-8$_7$2

M. T. Madsen, K. S. Berbaum, K. M. Schartz, and R. T. Caldwell, "Improved implementation of the abnormality manipulation software tools," *Proc. SPIE*, vol. 7966, Mar. 2011, pp. 7966121–7966127.

A. Pezeshk, N. Petrick, W. Chen, and B. Sahiner, "Seamless Lesion Insertion for Data Augmentation in CAD Training," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, 2017, pp. 1005–1015, doi: 10.1109/TMI.2016.2640180.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., WardeFarley, D., Ozair, S., Courville, A. C., Bengio, Y., 2014. Generative adversarial nets. In Conference on Neural Information Processing Systems, pages 2672–2680.

P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5967–5976.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. In IEEE Transactions on Biomedical Engineering.

Wolterink, J. M., Dinkla, A. M., Savenije, M. H. F., Seevinck, P. R., van den Berg, C. A. T., Isgum, I., 2017a. Deep MR to CT synthesis using unpaired data. In International Conference on Medical Image Computing and Computer Assisted Intervention. http://arxiv.org/abs/1708.01155.

K. Armanious, C. Jiang, M. Fischer, T. Kustner, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical Image Translation using GANs," *arXiv:1806.06397*, [Online]. Available: https://arxiv.org/pdf/1806.06397

Wolterink, J. M., Leiner, T., A. Viergever, M., Isgum, I., 2017b. Generative adversarial networks for noise reduction in low-dose CT. In IEEE Transactions on Medical Imaging.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., Wang, G., 2018c. Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. In IEEE Transactions on Medical Imaging, volume 37, pages 1348–1357.

M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion

classification," 2018, *arXiv:1803.01229*. [Online]. Available: https://arxiv.org/abs/1803.01229

C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with GANs," 2018, *arXiv:1803.01229*. [Online]. Available: https://arxiv.org/abs/1809.01410

A. Bissoto, F. Perez, E. Valle, and S. Avila, "Skin lesion synthesis with generative adversarial networks," 2018, *arXiv:1902.03253*. [Online]. Available: https://arxiv.org/abs/1902.03253

F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting data with GANs to segment melanoma skin lesions," *Multimed. Tools Appl.*, vol. 79, no. 21–22, 2020, pp. 15575–15592

K. Abhishek and G. Hamarneh, "Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis,"2019, *arXiv:1906.05845*. [Online]. Available: https://arxiv.org/abs/1906.05845

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, 2016.

S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, vol. 36, 2017.

J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *Conference on Computer Vision and Pattern Recognition*, 2018.

K. Armanious, Y. Mecky, S. Gatidis, and B. Yang, "Adversarial inpainting of medical image modalities," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3267–3271.

E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling gans for data augmentation in mammogram classification," in Image Analysis for Moving Organ, Breast, and Thoracic Images, pp. 98– 106. Springer, 2018

D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation," 2018, *arXiv:1806.04051*. [Online]. Available: arXiv:1806.04051

S. Liu, E. Gibson, S. Grbic, Z. Xu, A.A.A. Setio, J. Yang, B. Georgescu, and D. Comaniciu, "Decompose to manipulate: Manipulable Object Synthesis in 3D Medical Images with Structured Image Decomposition," 2018, *arXiv:1806.04051*. 2018, [Online]. Available: arXiv:1812.01737

Yang, J., Liu, S., Grbic, S., Setio, A. A. A., Xu, Z., Gibson, E., Chabin, G., Georgescu, B., Laine, A. F., Comaniciu, D. "Class-Aware Adversarial Lung Nodule Synthesis in CT Images," 2018, *arXiv:1812.11204*. [Online]. Available: arXiv:1812.11204

Xu, Z., Wang, X., Shin, H.-C., Yang, D., Roth, H., Milletari, F., Zhang, L., Xu, D. "Correlation via synthesis: end-to-end nodule image generation and radiogenomic map learning based on generative adversarial network," 2019, *arXiv:1812.11204*. [Online]. Available: arXiv:1812.11204

J. Ma, J. Chen, M. Ng, R. Huang, Y. Li,C. Li, X. Yang and Anne L. Martel, "Loss Odyssey in Medical Image Segmentation," Medical Image Analysis, 2021. https://doi.org/10.1016/j.media.2021.102035

T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2017, *arXiv:1708.02002*, [Online]. Available: https://arxiv.org/abs/1708.02002.

S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” 2017, *arXiv:1706.05721*, [Online]. Available: https://arxiv.org/abs/1706.05721.

N. Abraham, and N. M. Khan , "A novel focal tversky loss function with improved attention u-net for lesion segmentation," 2018, *arXiv:1810.07842*, [Online]. Available: https://arxiv.org/abs/1810.07842

C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” 2017, *arXiv:1707.03237*, [Online]. Available: https://arxiv.org/abs/1810.07842

B. Shirokikh *et al.* "Universal Loss Reweighting to Balance Lesion Size Inequality in 3D Medical Image Segmentation," 2020, *arXiv:2007.10033*, [Online]. Available: https://arxiv.org/abs/2007.10033

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, 2015, pp. 448–456.

D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd Int. Conf. Learn. Represent. ICLR 2015*, 2015, pp. 1–15.

L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern recognition* vol. 42 (9), 2009, pp. 1977-1987.

T. Perrot *et al.*, "Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning," *Eur Radiol* vol. 29, 4776–4782, 2019. https://doi.org/10.1007/s00330-019-6004-7

J. Jendeberg, P. Thunberg, M. Lidén, "Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network," *Urolithiasis.*, 49(1):41-49, 2021 doi:10.1007/s00240-020-01180-z

Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2003, pp. 1398–1402.